DOCTORAL THESIS

Vector Representations of Idioms in Data-Driven Chatbots for Robust Assistance

Tosin Adewumi

Machine Learning

LULEÅ UNIVERSITY OF TECHNOLOGY

Vector Representations of Idioms in Data-Driven Chatbots for Robust Assistance

Tosin Adewumi

Dept. of Computer Science, Electrical and Space Engineering
Luleå University of Technology
Luleå, Sweden

Supervisors:

Marcus Liwicki, Foteini Liwicki

There's no predicting the future.
Ironically though, I earn my living by making predictions. There's no predicting the future that my thesis would be what it is and I would be where I am today.
I dedicate this work to the ***All in all!***

ABSTRACT

This thesis presents resources capable of enhancing solutions of some Natural Language Processing (NLP) tasks, demonstrates the learning of abstractions by deep models through cross-lingual transferability, and shows how deep learning models trained on idioms can enhance open-domain conversational systems. The challenges of open-domain conversational systems are many and include bland repetitive utterances, lack of utterance diversity, lack of training data for low-resource languages, shallow world-knowledge and non-empathetic responses, among others. These challenges contribute to the non-human-like utterances that open-domain conversational systems suffer from. They, hence, have motivated the active research in Natural Language Understanding (NLU) and Natural Language Generation (NLG), considering the very important role conversations (or dialogues) play in human lives.

The methodology employed in this thesis involves an iterative set of scientific methods. First, it conducts a systematic literature review to identify the state-of-the-art (SoTA) and gaps, such as the challenges mentioned earlier, in current research. Subsequently, it follows the seven stages of the Machine Learning (ML) life-cycle, which are data gathering (or acquisition), data preparation, model selection, training, evaluation with hyperparameter tuning, prediction and model deployment.

For data acquisition, relevant datasets are acquired or created, using benchmark datasets as references, and their data statements are included. Specific contributions of this thesis are the creation of the Swedish analogy test set for evaluating word embeddings and the Potential Idiomatic Expression (PIE)-English idioms corpus for training models in idiom identification and classification. In order to create a benchmark, this thesis performs human evaluation on the generated predictions of some SoTA ML models, including DialoGPT. As different individuals may not agree on all the predictions, the Inter-Annotator Agreement (IAA) is measured. A typical method for measuring IAA is Fleiss Kappa, however, it has a number of shortcomings, including high sensitivity to the number of categories being evaluated. Therefore, this thesis introduces the Credibility unanimous score (CUS), which is more intuitive, easier to calculate and seemingly less sensitive to changes in the number of categories being evaluated. The results of human evaluation and comments from evaluators provide valuable feedback on the existing challenges within the models. These create the opportunity for addressing such challenges in future work.

The experiments in this thesis test two hypotheses; 1) an open-domain conversational system that is idiom-aware generates more fitting responses to prompts containing idioms, and 2) deep monolingual models learn some abstractions that generalise across languages. To investigate the first hypothesis, this thesis trains English models on the PIE-English idioms corpus for classification and generation. For the second hypothesis, it explores cross-lingual transferability from English models to Swedish, Yorùbá, Swahili, Wolof, Hausa, Nigerian Pidgin English and Kinyarwanda. From the results, the thesis' additional contributions mainly lie in 1) confirmation of the hypothesis that an open-domain conversational system that is idiom-aware generates more fitting responses to prompts containing idioms, 2) confirmation of the hypothesis that deep monolingual models learn some abstractions that generalise across languages, 3) introduction of CUS and its benefits, 4) insight into the energy-saving and time-saving benefits of more optimal embeddings from relatively smaller corpora, and 5) provision of public access to the model checkpoints that were developed from this work. We further discuss the ethical issues involved in developing robust, open-domain conversational systems. Parts of this thesis are already published in the form of peer-reviewed journal and conference articles.

Acknowledgements

My deepest gratitude goes to all who have supported my PhD journey in the Machine Learning Group, the Embedded Intelligent Systems Lab (EISLAB), and the Department of Computer Science, Electrical and Space Engineering at Luleå University of Technology. Particularly, my profound appreciation goes to Professor Marcus Liwicki, my supervisor, who advised me, before I was even hired, to *"find and join networks"* to foster my work. This advice led me to Masakhane - the African network of over 1,000 NLP-related researchers. My wholehearted gratitude goes to Assistant Professor Foteini Liwicki, my assistant supervisor, who advised me shortly after I was hired to *"write a SotA paper"*. I'm thankful to many people, too numerous to list here, including Professor Jonas Ekman (the head of the department), Ulf Bodin, Petter Kyösti, Björn Backe, all the seniors, the course instructors I had the opportunity of learning from, the administrative staff (particularly Karin Rosengren), and the members of the examining committee (Johan Boye, Taiwo Kolajo, Viggo Kann, Andre Freitas, and Diana Chronéer).

I certainly cannot forget the overwhelming support of my dad, mum, siblings, nieces, nephews, and friends (including Seye Olumide). I have been influenced one way or the other by those I have met or worked with, including the lunch-hour mates, Hamam, who wears a smile always, and other colleagues who have turned life-time friends for me. Below, I capture some interesting words from a few of them. Thank you all for making me laugh and adding warmth to the cold weather of Luleå. Again, there would be no me (or this work) without the All in all; I'm grateful.

"Whatever journey I need to make, I'll have a companion." - Bukky Peters

"A meeting doesn't have to last an hour; if it takes two minutes, that's enough." -Lama Alkhaled

"Good things will come at the right time..don't push it. Just work hard and believe." - Sana Al-Azzawi

"You are the hero of your own story." -Nosheen Abid

"I think having a flower to grow makes your life happier." -Maryam Pahlavan

"Yaaaay! This is the best Christmas ever." -Monife Onamusi (6 years old)

Luleå, June 2022
Tosin Adewumi

Contents

Publications

Publications included

1. Adewumi, T., Brännvall, R., Abid, N., Pahlavan, M., Sabry, S. S., Liwicki, F., & Liwicki, M. (2021). Småprat: DialoGPT for Natural Language Generation of Swedish Dialogue by Transfer Learning, *Proceedings of the Northern Lights Deep Learning Workshop 2022*, 3, doi.org/10.7557/18.6231

2. Adewumi, T., Vadoodi, R., Tripathy, A., Nikolaidou, K., Liwicki, F., & Liwicki, M. (2022). Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms, *International Conference on Language Resources and Evaluation (LREC)*, 13, (Accepted).

3. Adewumi, Tosin, Liwicki, Foteini and Liwicki, Marcus. "Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks" *Open Computer Science*, 12(1), 2022, pp. 134-141. https://doi.org/10.1515/comp-2022-0236

4. Adewumi, T.P.; Liwicki, F.; Liwicki, M. Conversational Systems in Machine Learning from the Point of View of the Philosophy of Science—Using Alime Chat and Related Studies. *Philosophies* 2019, 4(41), https://doi.org/10.3390/philosophies4030041

5. Adewumi, T. P., Liwicki, F., & Liwicki, M. (2020). Corpora compared: The case of the swedish gigaword & wikipedia corpora, *The Swedish Language Technology Conference*, 8.

6. Adewumi, T. P., Liwicki, F., & Liwicki, M. (2020). The Challenge of Diacritics in Yoruba Embeddings, *ML4D Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS).*

7. David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin

Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, Salomey Osei; MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics* 2021; 9 1116–1131. doi: https://doi.org/10.1162/tacl_a_00416

8. Adewumi, Tosin; Adeyemi, Mofetoluwa; Anuoluwapo, Aremu; Peters, Bukola; Buzaaba, Happy; Samuel, Oyerinde; Rufai, Amina Mardiyyah; Ajibade, Benjamin; Gwadabe, Tajudeen; Traore, Mory Moussou Koulibaly; Ajayi, Tunde; Muhammad, Shamsuddeen; Baruwa, Ahmed; Owoicho, Paul; Ogunremi, Tolulope; Ngigi, Phylis; Ahia, Orevaoghene; Nasir, Ruqayya; Liwicki, Foteini; Liwicki, Marcus (2022). Ìtàkúròso: Exploiting Cross-Lingual Transferability for Natural Language Generation of Dialogues in Low-Resource, African Languages *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Submitted).

9. Adewumi, T. P., Liwicki, F., & Liwicki, M. (2022). Vector Representation of Idioms in Conversational Systems (2022) *Information*, MDPI (Submitted).

10. Adewumi, T. P., Liwicki, F., & Liwicki, M. (2022). Exploring Swedish & English fastText embeddings, *International Workshop on Artificial Intelligence and Cognition (AIC)*, (Submitted).

Other publications

1. Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, et al.. (2021). The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. *In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics

2. Adewumi, T., Alkhaled, L., Alkhaled, H., Liwicki, F., & Liwicki, M. (2022). ML_LTU at SemEval-2022 Task 4: T5 Towards Identifying Patronizing and Condescending Language. *International Workshop on Semantic Evaluation* (Accepted).

3. Adewumi, T. P., & Liwicki, M. (2020). Inner for-loop for speeding up blockchain mining. *Open Computer Science*, 10(1), 42-47.

4. Sabry, S. S., Adewumi, T., Abid, N., Kovacs, G., Liwicki, F., & Liwicki, M. (2022). HaT5: Hate Language Identification using Text-to-Text Transfer Transformer, *International Joint Conference on Neural Networks (IJCNN)*, (Accepted).
5. Javed, Saleha, Tosin P. Adewumi, Foteini S. Liwicki, and Marcus Liwicki. (2021). "Understanding the Role of Objectivity in Machine Learning and Research Evaluation" *Philosophies* 6, no. 1: 22. https://doi.org/10.3390/philosophies6010022
6. Saini, R., Kovács, G., Faridghasemnia, M., Mokayed, H., Adewumi, O., Alonso, P., & Liwicki, M. (2021). Pedagogical Principles in the Online Teaching of Text Mining: A Retrospection. *In Proceedings of the Fifth Workshop on Teaching NLP* (pp. 1-12). Association for Computational Linguistics
7. Adewumi, T., Sabry, S. S., Abid, N., Liwicki, F., & Liwicki, M. (2022) The Love Model: Hate Speech, Data Augmentation & Ensemble (Submitted).
8. Adewumi, T., Liwicki, F., & Liwicki, M. (2022). State-of-the-art in Open-domain Conversational AI: A survey (Submitted).
9. Adelani, D., et al. (2022) Choosing the Best Transfer Language for Named Entity Recognition: A Study on African languages (Submitted).
10. Gehrmann, S., et al. (2022) GEMv2: NLG benchmarking in 45 languages with a single line of code (Submitted)

Chapter 1

Introduction

*"It is **the beginning of the end** of the bland chatterbox."*

(Paradox)

A major measure of human intelligence is the ability to communicate in natural language (Adiwardana et al., 2020). The more colourful[1] the language of expression, the more culturally rich a society may be counted to be. NLP is the study of the modes of human language for scientific purposes. It is an intersection of the fields of linguistics and computer science (Jurafsky and Martin, 2020). Some of the main goals of NLP are to understand and generate natural language from data (Jurafsky and Martin, 2020). The increasingly dominant approach to achieve these goals is to use neural NLP, which has succeeded statistical NLP (Zhou et al., 2020b). Statistical NLP purely uses information from a training dataset to establish possible events, such as which characters are most likely to form words (Indurkhya and Damerau, 2010) while neural NLP is centred on using artificial neural network (ANN), in addition to data, for the goals and tasks of NLP. NLP itself is a part of Machine Learning (ML), which, according to Mitchell et al. (1997), is the use of a program, say M, to possibly learn from experience E with regards to a task or class of tasks T and performance metric P, so that the performance at tasks in T, as measured by P, improves with experience E (Hackeling, 2017).

This chapter gives a gentle introduction to some of the concepts, philosophy, and the scientific method this work uses. The chapter introduces conversational systems and the ways of evaluating them, especially using some version of the Turing test. In addition, it discusses the benefits and challenges of conversational systems and the contributions of this work. The chapter concludes with ethical considerations when conducting research generally, but specifically for conversational systems, and highlights some related work in the field.

[1] colourful here means "rich" - dictionary.com

1.1 Background

Historically, work in NLP began as soon as the early days of the computer (Jurafsky and Martin, 2020). Some notable contributions came from the work of Turing et al. (1936), the work of McCulloch and Pitts (1943) on the neuron, Kleene et al. (1956), and Chomsky (1956). Their early work birthed the field of formal language theory. A formal language consists of sequences of symbols or words that are well-formed according to a specific set of rules (Jurafsky and Martin, 2020). They can be defined using set theory or algebra (Chomsky, 1956). Shannon's contribution gave rise to the development of probabilistic models to automata for language (Shannon, 1948). The development witnessed in speech recognition in those early periods came about through the stochastic approach (Jurafsky and Martin, 2020). The return of empiricism around the 1980s and early 1990s witnessed the rise of probabilistic methods, increasing use of data-driven techniques for various NLP tasks, new direction on model evaluation by using held-out data, emphasis on comparison of performance with previously published work, and increased volume of work on NLG.

Chatbots are systems with the ability to mimick the unstructured conversations that are typical of human-human chats by communicating in natural language with users (Jurafsky and Martin, 2020). They can be designed for different purposes, such as making task-oriented agents more natural or for entertainment. Chatbots, conversational systems and dialogue systems are used interchangeably in this work. A chatbot may be designed as a simple rule-based template system or may involve more complex ANN architectures that are trained on large datasets to generate responses. The first acclaimed conversational system was ELIZA (Weizenbaum, 1969). The example conversations of the system, as demonstrated by Weizenbaum (1969), show how therapeutic the responses can be. People reportedly became so engrossed with the program and were possibly having private conversations with it (Jurafsky and Martin, 2020). Some modern systems are still architectured in the rule-based fashion of ELIZA (Jurafsky and Martin, 2020). An example is PARRY (Colby et al., 1971). Besides having a regular chat, conversational systems can be designed to express emotions. PARRY, for example, was designed to express fear and anger, depending on the topic of conversation (Colby et al., 1971). The method of evaluating conversational systems varies, depending on the type of system at hand. For open-domain conversational systems, human evaluation of how human-like the responses or conversations are is usually common (Zhang et al., 2020). This type of evaluation usually resembles the Turing test format.

1.1.1 The Turing test

The Turing test (or indistinguishability test) is possibly the ultimate test of human-like conversation such that a human is not able to distinguish if the responses or conversations are from another human or a machine. Two systems, S_a and S_b, are input-output equivalent in a particular scenario, when their input-output pairs are not distinguishable in respect to specified dimensions (Colby et al., 1971). It is important to note that the output for our reference system for a given input, in many cases, is actually a set of possible candidate outputs. These candidate outputs are referred to as the reference for evaluating the performance of NLG systems (the imitation) for some metrics, such as the BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) score. More is discussed about such metrics in Section 4.4.

Turing (1950) proposed, originally, to consider the question "Can machines think?", which some considered baseless. He replaced such a formulation with a relatively unambiguous one, which is designed as the 'imitation game'. The reformulated question is "Are there imaginable digital computers which would do well in the imitation game?" (Turing, 1950). A man, a woman, and an interrogator of either sex, who is in a separate room from the man and the woman, are players of the game. The objective for the interrogator is to determine who is the man and who is the woman. The interrogator does this by posing questions to the man and woman, which are answered in some written format. The objective of the man is to trick the interrogator into believing he's a woman while the objective of the woman is to convince the interrogator she's a woman. When a machine (or digital computer) replaces the man, the test seeks to know if the interrogator will decide wrongly as often as when it was played with a man (Turing, 1950). Figure 1.1 depicts the 'imitation game' for Man/Woman (top) and Machine/Woman (bottom).

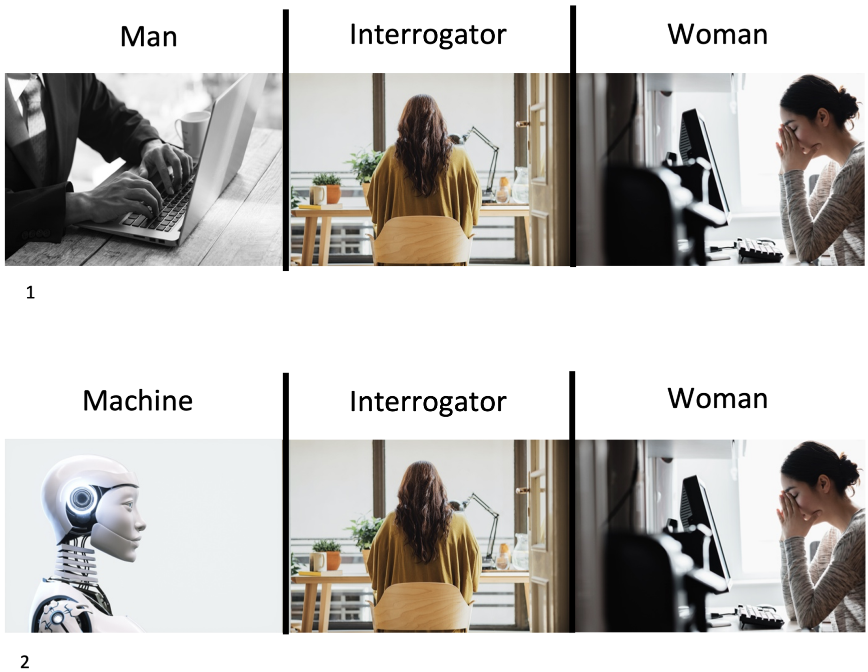

Figure 1.1: Depiction of the Turing test (The 'imitation game')

One should note that there are objections to the concept of a machine thinking (Colby et al., 1972; Shieber, 1994; Turing, 1950). They include the incompleteness theorem, which argues that there are limits to questions that a machine based on logic can answer (Gödel, 1931; Turing, 1950). Also, the assertion that the analytical engine does not presume to originate anything by Ada Lovelace[2] (Fuegi and Francis, 2003) is viewed as a strong objection (Turing, 1950). Other objections include the theological objection, which he found fault with; the 'heads in the sand' objection, which dreads the consequences of machines being able to think but for which Turing offers consolation; the argument from consciousness, which emphasises thoughts and emotions as what should be the source of the machines ability (Turing, 1950). The Turing test has different versions (Traiger, 2003). Indeed, at some point in the same paper by Turing (1950), after replacing the man with a machine, the woman is also replaced by a man. Turing's formulation of the imitation game does not precisely match modern versions of the test (Saygin and Cicekli, 2002). Despite the objections to the main question of machines thinking, the fact that the Turing test provides a means to measure performance is a good thing.

This test was applied to PARRY, a chatbot designed to imitate aggressive emotions, like a paranoid person (Colby et al., 1972). Most psychiatrists (23 out of 25) couldn't distinguish between text transcripts of PARRY and real paranoids, so it is the first system to pass this test, at least, the early version of the test (Colby et al., 1971; Jurafsky and Martin, 2020). However, this is disputed by some, since ELIZA was able to fool many of its users as well (Mauldin, 1994; Jurafsky and Martin, 2020). Also, the example of PARRY can be argued to be an edge case since the comparison was made with paranoids instead of rational human beings (Mauldin, 1994). A restricted version of the Turing test was introduced in 1991, alongside the unrestricted version, in what is called the Loebner Prize competition (Mauldin, 1994). Prizes have been awarded every year to conversational systems that pass the restricted version of the competition (Bradeško and Mladenić, 2012). The Loebner Prize competition has its share of criticisms. It is viewed as rewarding tricks instead of furthering the course of AI (Shieber, 1994; Mauldin, 1994). Shieber (1994) recommended an alternative approach that would involve a different award methodology, which is based on a different set of assessment, that is done on an occasional basis.

1.1.2 Assumptions

Certain assumptions are essential when solving certain tasks (Elkner et al., 2010). Adewumi et al. (2019) argue that, in line with the assumptions alluded to by Kuhn (1970), the scientific community holds on to some assumptions about our world. These assumptions are essential for us to understand the way the world works and how we perceive things. We approach this work from a Naturalist philosophical point of view (Creath, 2011; Javed et al., 2021). Central to the Naturalist philosophical point of view are a collection of beliefs and values, which are untested by the scientific processes but give legitimacy to the scientific systems. They also set the boundaries of investigations. The type of as-

[2]fourmilab.ch/babbage/sketch.html

sumptions we refer to are stable and not the quickly-evolving postulations that Longino (2020) describe as lacking in objectivity. In the field of NLP some of the assumptions we make are identified below:

- Random sampling is representative for an entire population (Kazmier, 2004).
- The probability distribution of samples from a population follow the normal distribution, for a minimum sample size of 30. This is based on the central limit theorem (Kwak and Kim, 2017)
- Idioms are often language specific (Alm-Arvius, 2003). This implies many idioms have unique meanings within the cultural language they evolve in.
- Language processing is incremental. (Clark et al., 2012). This implies each newly encountered word is integrated immediately into the interpretation of what has been read.
- Models use left to right decomposition of the text probability to compute the probability of generating a complete sequence (Holtzman et al., 2020). It should be noted that there are languages that function from right to left. Examples include Hebrew and Arabic.

1.1.3 Natural Language Processing (NLP) Tasks

There are many tasks within NLP, including downstream tasks (Gatt and Krahmer, 2018; Gehrmann et al., 2021). Downstream tasks are the end-tasks of importance to users of NLP systems (Gatt and Krahmer, 2018). NLP tasks are focused around NLU, NLG, and other auxiliary tasks that support the former two areas. Some NLP tasks are briefly discussed below.

- Text Classification (TC) is a general term for the many types of classification tasks that exist in NLP. It mainly involves categorising tokens of sequences or blocks of text, in what may also be document categorisation (Kowsari et al., 2019), into the different categories that may be defined (Aggarwal and Zhai, 2012). Classification variants that exist include: binary, multiclass, multilabel, open-class (where the labels are not defined in advance), and sequence classification (where a set of inputs are jointly classified) (Bird et al., 2009). Examples of specific TC include Sentiment Analysis (SA), hate speech (Sabry et al., 2022), and Patronising and Condescending Language (PCL) (Pérez-Almendros et al., 2022; Adewumi et al., 2022b).
- Named Entity Recognition (NER) involves the classification of specific entities. It's a task of sequence tagging that is useful in Information Retrieval (IR), conversational systems, and other applications (Adewumi et al., 2022d; Adelani et al., 2021).

- Sentiment Analysis (SA) is a type of TC that involves classification of sentences/text according to sentiments or opinion (Aggarwal and Zhai, 2012; Medhat et al., 2014; Zhang et al., 2018a).

- Text Summarisation involves summarising relevant points within a large text. Summarisation requires NLP systems to generate human-readable summaries of long sequences of text (Aggarwal and Zhai, 2012; Gatt and Krahmer, 2018).

- Machine Translation (MT) involves translating text from one language to a second, target language (Vaswani et al., 2017). The use of parallel corpora is common for this task. Large quantities of parallel texts (or corpora) from news and government website that publish in multiple languages are often used. Before feeding a model, text alignment may be carried out to pair up sentences, given a pair of documents in two languages (Bird et al., 2009). N-gram-based automatic metrics are the dominant metrics for evaluating MT systems (Sammons et al., 2012).

- Recognizing Textual Entailment (RTE) focuses on general text inference capabilities (Sammons et al., 2012). It is an NLU task where systems are required to find evidence to support a hypothesis (Bird et al., 2009). It has the potential to benefit other NLP tasks. A sequence of text entails a hypothesis if the meaning of the hypothesis can be deduced from the meaning of the text sequence (Sammons et al., 2012). It is a directional relationship between the pair of texts. The point is whether conclusion can be drawn that a piece of text contains reasonable evidence for describing a hypothesis to be true, as a human would, rather than based on logical entailment (Bird et al., 2009). Since there's the existing challenge with systems not being able to reason, a key objective in NLP research is to understand language by using strong techniques instead of unrestricted knowledge or reasoning capabilities (Bird et al., 2009). Lexical matching is probably the simplest way of solving the task of RTE but this approach is too simplistic for more challenging situations.

- Word Sense Disambiguation (WSD) finds the intended sense of a word within a context. One way of identifying what a pronoun or noun refers to in a sentence is through anaphora (pronoun) resolution. Semantic role labeling is another technique, which identifies how a noun phrase relates to the verb (as agent, patient, etc) (Bird et al., 2009).

- Information Retrieval (IR), which is a more general case of information extraction, recognises instances of a fixed set of relations in a set of documents (Sammons et al., 2012).

- Question Answering (QA) requires NLP systems to deduce candidate answers to a question from areas of a fixed document (Sammons et al., 2012).

- Question Generation (QG) involves a system generating a relevant question from a block of text, such as sentences or paragraphs (Rus et al., 2011).

- Co-reference resolution involves settling if an entity mentioned in one place refers to another entity mentioned in another place within a given sequence of text (Sammons et al., 2012).

- Natural Language Generation (NLG), which is the main focus of this thesis, comprises some of the above-mentioned tasks (QA, QG) and some additional tasks focusing on generating text from text or other kinds of data (Gatt and Krahmer, 2018; Gehrmann et al., 2021; Reiter and Dale, 1997, 2000). These tasks are usually base on three stages, as shown in Figure 1.2: document planning, microplanning, and realisation. Those stages are further divided into the following sub-stages (Reiter and Dale, 2000; Reiter, 2010)

  – Content Determination - this involves determining the information to be communicated.
  – Text Structuring - this involves determining the order of presentation of texts.
  – Lexical choice - this involves determining words or phrases for expression.
  – Referring Expression - this involves selecting words to identify entities within a domain.
  – Syntactic choice - this determines the syntax construction
  – Aggregation - this involves grouping of related messages.
  – Overgeneration - this involves generating the right morphological forms.
  – Selection - selecting the most probable text from the generated set.

1.1.4 Natural Language Generation (NLG) and conversational systems

Human conversation can be complicated, though we may take them for granted because we are accustomed to them. Section 4.1 describes some of the characteristics of human conversation. Making conversational systems learn the intricacies of side sequence (or sub-dialogue) within a main dialogue (Jefferson, 1972), clarification question or presequences (before a main request) can be a challenging effort. Furthermore, in natural conversations, initiative can shift between two speakers and this is a challenge in conversational systems, as they are usually designed to be passive responders (Jurafsky and Martin, 2020).

Of the various architectures for conversational systems, frame-based architecture (or Genial Understander System (GUS)) is common with task-oriented systems (Bobrow et al., 1977) while rule-based and data-driven architectures are the common architectures with open-domain systems. Section 4.2 discusses more about this. A modern, sophisticated frame-based architecture is called a dialogue-state (Jurafsky and Martin, 2020). The GUS architecture for frame-based dialog system is used in Siri, Alexa, and Google

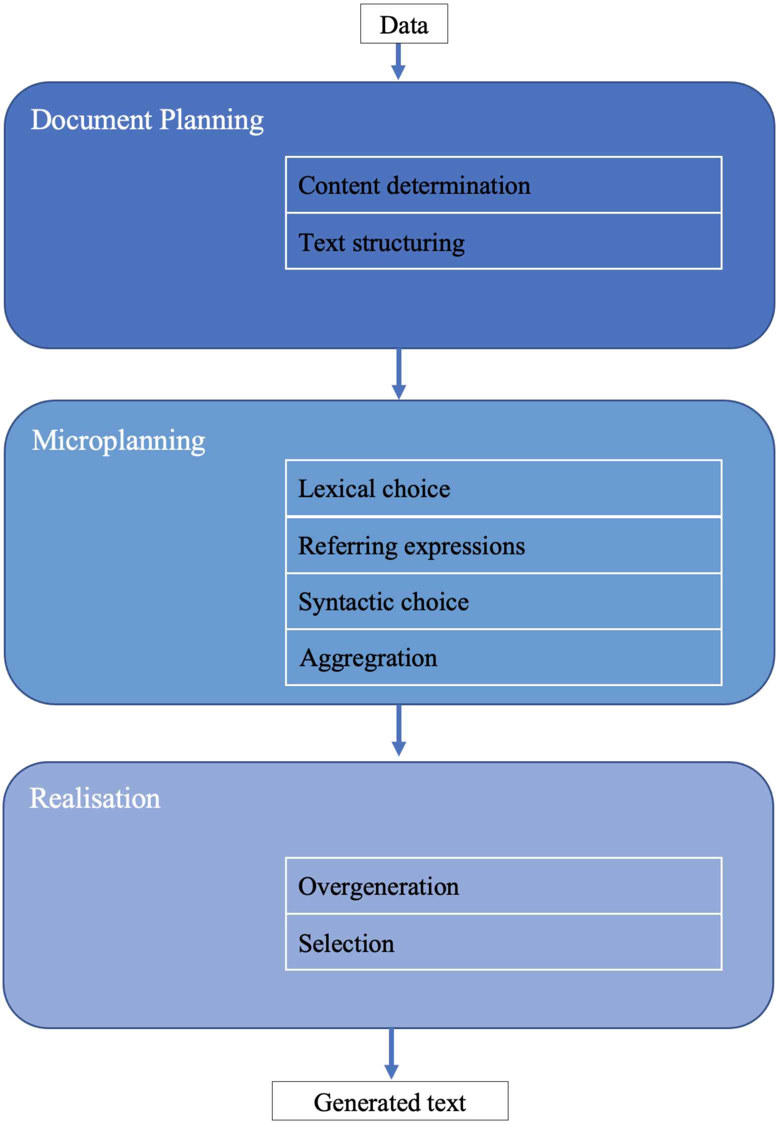

Figure 1.2: Depiction of the NLG pipeline, based on Reiter (2010)

Assistant. It is a production rule system because different types of inputs cause different productions to fire. It also has condition-action rules attached to slots (Chowdhary, 2020). Frame-based system's language generation module uses template-based generation, where all or most of the utterances to the user are hand-crafted by the dialogue designer (Chowdhary, 2020).

Examples of data-driven architectures include Information Retrieval and encoder-decoder architectures. Data-driven conversational systems are data-intensive, as they require a lot of data for training the system (Jurafsky and Martin, 2020). One approach that has gained popularity in usage is to pretrain on large datasets of text or conversations from Reddit, Twitter or other social media data before finetuning on a specific dataset (Jurafsky and Martin, 2020; Zhang et al., 2020). Examples of NLG systems include SumTime, which involves weather forecast, and SkillSum, which involves educational assessment (Reiter, 2010). Such can be extended to have continual output based on user input in order to have a conversation with the user. In Figure 1.2, the final realisation stage seems to be the most understood part of the pipeline and probably receives the most attention in terms of implementation (Reiter, 2010). It is noteworthy that not all the stages of the pipeline are used in all NLG systems.

1.2 Benefits of conversational systems

The marginal benefit or value a thing holds over the possible risks usually determines whether it is worthwhile to pursue investments in such a thing. Research in conversational systems have been growing since the early days of ELIZA because of the apparent benefits (Jurafsky and Martin, 2020). These benefits have led to huge investments in conversational systems technology by many organisations. Some of those benefits are highlighted below.

- The provision of psychological or psychiatric treatment for humans based on favourable behaviour determined from experiments designed to modify input-output behaviour in models (Colby et al., 1971).
- The support of users that have disabilities, such as blindness (Reiter, 2010).
- The seamless accomplishment of specific tasks, such as airline bookings and hotel reservations (Jurafsky and Martin, 2020).
- Provision of therapeutic company.
- Conduit of world/domain knowledge (Reiter, 2010).
- Provision of educational content in a concise mode (Kerry et al., 2008).
- Automated generation of quality data for low-resource languages (Adewumi et al., 2022a).

1.3 The challenges of open-domain conversational systems

The road to a human-like conversational AI system is fraught with challenges. These challenges contribute to the non-human-like utterances which open-domain conversational systems tend to have but they also motivate active research in NLP, considering the very important role conversations play in our lives. Progress has been noticeable in some areas, however, other areas have witnessed little advancement. Some of the challenges are highlighted below. The last three items in the list form part of the important research questions in this work and are discussed a little more in the next section.

- Lack of coherence in sequence of text or across multiple turns of generated turns of conversation (Jurafsky and Martin, 2020; Welleck et al., 2019).
- Non-empathetic responses from conversational systems (Rashkin et al., 2019).
- Lack of utterance diversity (Holtzman et al., 2020).
- Lack of memory to personalise user experiences.
- Bland repetitive utterances (Holtzman et al., 2020)
- Initiative coordination (Jurafsky and Martin, 2020)
- Poor inference and implicature during conversation.
- Lack of training data for low-resource languages (Adewumi et al., 2020a).
- Shallow world-knowledge in conversational systems.
- Developing ethical and robust conversational systems.
- Utilising figures of speech (idioms) in models to enhance NLP.
- Gaining robust assistance or performance from models trained on figures of speech (idioms) to enhance open-domain conversational systems.

1.4 Research questions

The main goal of this thesis is to generate conversations that are more fitting for contexts where idioms are present. After conducting a systematic literature review and identifying gaps, the following four research questions (RQ) arose. Addressing these questions to some meaningful point will contribute to the furtherance of open-domain conversational systems, some of which are mentioned in Section 1.5. The general approach that is used to address these questions is described in Section 1.8.

RQ1 How importantly do hyper-parameters influence word embeddings' performance?

RQ2 What factors are important for developing ethical and robust conversational systems?

RQ3 To what extent can models trained on figures of speech (idioms) enhance NLP?

RQ4 How can models trained on figures of speech (idioms) enhance open-domain, data-driven chatbots for robust assistance?

1.5 Hypotheses and contributions

This work investigates the four RQs mentioned earlier. It tests the following two hypotheses (H):

H1 An open-domain conversational system that is idiom-aware generates more fitting responses to prompts containing idioms. This is investigated in controlled experiments by comparing similar models whereby one is exposed by training to a dedicated idioms data (in this case, the PIE-English corpus) and the other is not.

H2 Deep monolingual models learn some abstractions that generalise across languages (Artetxe et al., 2020). This is investigated by exploring cross-lingual transferability for seven languages from English models to Swedish, Yorùbá, Swahili, Wolof, Hausa, Nigerian Pidgin English, and Kinyarwanda, most of which are low-resource languages.

As a result of the conclusions from various empirical studies carried out, the following are the contributions of this thesis.

1. We created and publicly provide, under the Creative Commons Attribution 4.0 (CC-BY4) licence, the Swedish analogy test set for evaluating Swedish word embeddings (Adewumi et al., 2020b). This addresses RQ1. The resource was verified by Språkbanken and is hosted on the Swedish Språkbanken website[3].

2. We created and publicly provide the Potential Idiomatic Expression (PIE)-English idioms corpus, under the CC-BY4 licence, for training models in idiom identification and classification (Adewumi et al., 2021). This addresses RQ3 and RQ4. The resource is hosted on the International Conference on Language Resources and Evaluation (LREC) platform[4].

3. We created and publicly provide the AfriWOZ dialogue dataset of parallel corpora of 6 African languages under the CC-BY4 licence, primarily for training open-domain conversational systems (Adewumi et al., 2022a). The dataset may be adapted for other relevant NLP tasks, like MT. This addresses RQ2. The resource is hosted online[5].

[3]spraakbanken.gu.se/en/resources/analogy
[4]lrec2022.lrec-conf.org/en/
[5]github.com/masakhane-io/chatbots-african-languages

4. We confirm the hypothesis that an open-domain conversational system that is idiom-aware generates more fitting responses to prompts containing idioms. We make the conversational models idiom-aware by training on the PIE-English idioms corpus. This, therefore, enhances open-domain conversational systems and addresses RQ3 and RQ4.

5. We confirm the hypothesis that deep monolingual models (in this case, English) learn some abstractions that generalise across languages (Adewumi et al., 2022c,a). This contributes to addressing RQ2. We show from human evaluations of the transcripts of the conversational models that six out of the seven target languages are transferable to. The only language that seems not transferable to, in a conversational setup, is the Yorùbá language. To the best of our knowledge, this work may be the first work exploring crosslingual transferability from deep monolingual English models to low-resource languages for open-domain conversational systems.

6. We introduce the Credibility unanimous score (CUS). This is an Inter-Annotator Agreement (IAA) metric that is based on homogeneous samples in the transcript or data for which IAA is to be determined. It contributes to addressing RQ2. The score is based on the simple percentage of the unanimous votes of the annotators over the homogeneous samples. The homogeneous samples serve two additional purposes, besides providing a basis for IAA. These are 1) to test the credibility of the annotators, and 2) to determine majority agreement on the transcript; in this case, agreement on human-human conversations.

7. We provide insight into the energy-saving and time-saving benefits of more optimal embeddings from better hyperparameter combinations and relatively smaller corpora (Adewumi et al., 2022d). This addresses RQ1 and also contributes to RQ2.

8. We created and publicly provide access to a selected set of word embeddings in English, Swedish and Yorùbá (Adewumi et al., 2022d, 2020a,b).

9. We open-source all the codes used in this work and host them on Github[6], under the CC-BY4 licence. It also contributes to addressing RQ2.

10. We provide public, free access to all the model checkpoints that were developed in the course of this work on the HuggingFace hub[7] (Adewumi et al., 2022c; Adelani et al., 2021; Adewumi et al., 2022a). This also contributes to addressing RQ2.

11. We develop the philosophical argument for developing robust and ethical conversational systems (Adewumi et al., 2019; Javed et al., 2021). It addresses RQ2. This may serve as a springboard for further helpful discussions around the subject.

[6] github.com/tosingithub
[7] huggingface.co/tosin

1.6 Basics of artificial neural network (ANN)

There are three components that describe an artificial neural network or model, according to Bird et al. (2009). These are the model's architecture or topology, the activation function, and the weights' learning algorithm. While this work does not focus on the mathematical exposition of ANN and other concepts, we provide brief plain descriptions. The number of neurons determine the number of parameters in an ANN, which determine the complexity of the network. An ANN may contain connected neurons at different depths. The NN is termed shallow when the depth is only a few layers (say, two or three). The objective with ANN is to find the weights which minimise the value of a cost function while approximating or solving a particular function (Hackeling, 2017). Information in the NN is processed collectively in parallel throughout a network of nodes (or neurons) and the output of the neuron is generated by passing its processed (or summed) inputs through an activation function (Shiffman et al., 2012).

Parameters refer to weights, bias, and other properties of an NN, which are trained by some optimisation method. A neuron requires the additional input, called bias, which has a constant value of 1 or some other constant. This helps to avoid null processed input from the original inputs (Shiffman et al., 2012). The cost function is also called the loss function and it is used to define and measure the error of a model. Training or test errors are differences between the prediction and observed values of the training data or test data, respectively (Hackeling, 2017). If the number of neurons in a neural net is too large, it will likely overfit the training data. Unlimited data makes overfitting unlikely. The problem of overfitting implies the network is not able to know the true function in the regions where there is no data, making it an error of interpolation (Bird et al., 2009). A model that memorises (by overfitting) the dataset may not perform well generally when tested. It is very likely to memorise structures that are noise within the data (Hackeling, 2017). The dev (or validation) set is used to tune hyperparameters, which control how models learn.

Prediction error may arise because of two main reasons: the bias of a model or its variance (Hackeling, 2017). Overfitting and underfitting occur in models with high variance and high bias, respectively. It is usually preferred to have bias-variance trade-off so that we have low bias and low variance. Unfortunately, efforts to keep one low increases the other (Hackeling, 2017). To reduce overfitting, some of the methods available are the following: early stopping, drop out, and regularisation. Early stopping is when we stop the training as soon as performance on the validation set starts to deteriorate, which will be apparent from a rising validation loss. Drop out implies a certain percentage of the neurons are dropped in the network; dropping out 20% of the input and 50% of the hidden units is usually found to be optimal, however, a disadvantage of dropout is that it may take two or three times longer to train (Srivastava et al., 2014). Regularisation, which is applied to reduce overfitting, is a collection of techniques for preventing overfitting (Hackeling, 2017). It penalises complexity, in line with the principle of parsimony (or Ockham's razo). The penalty could be *L1* or *L2* regularisation. The principle of parsimony suggests that entities need not be multiplied unnecessarily or a simpler model (with fewer parameters) should be preferred over a complex one for explaining observations. The use of the principle reduces the possibility of errors (Hagan et al., 1997). It, therefore, finds the simplest model that explains the data. Least absolute shrinkage and selection operator (LASSO) and ridge regression are special cases of regularisation techniques. In these, the hyperparameters for *L1* or *L2* penalty are set equal to zero. Hyperparameters, unlike model weights, are parameters that are not learned automatically during training but set manually, usually before training. They are user-tuned and examples are the number of neurons, layers, learning rate, regularisation penalty, momentum, number of epochs, batch size, dropout rate, etc (Hackeling, 2017).

Backpropagation is used to update model weights so that the model can learn how to map arbitrary inputs to outputs (Rumelhart et al., 1985; Clark et al., 2012). It is a gradient descent method for obtaining the weights that minimise the system's performance error (Rumelhart et al., 1985). It solves the problem of the analytical approach by estimating the optimal parameters. The analytical approach is undesireable, especially when there are hundreds of thousands of inputs, which create a computational menace of inverting the derived square matrix while trying to obtain the weights (Hackeling, 2017). Gradient descent is slow in practice and two main approaches to its implementation are heuristic techniques (such as learning rate variation) and standard numerical optimisation techniques (Hackeling, 2017). Their derivatives are used to update the weights of the model differently. The use of momentum implies application of a momentum filter to backpropagation by using a coefficient between 0 and 1. This helps to accelerate convergence of the algorithm as the trajectory moves in a consistent direction. There is more momentum in the trajectory when there is a larger momentum assigned (Hagan et al., 1997). It is important to point out that we may not be sure that the algorithm converges at an optimum solution, hence, it is best to try a number of different initial conditions in order to ensure that an optimum solution is obtained. The learning rate is a crucial hyperparameter of gradient descent. In addition, increasing the learning rate when the surfaces are flat but decreasing the rate when the slope increases will speed up convergence (Hagan et al., 1997).

ANN models may be trained as classifiers through supervised learning with annotated data. These may then be used to make predictions on unseen data (or test set). Typically, there are two types of model classifiers: generative model classifiers, which predict based on the joint probability of input-label pair, and conditional (discriminative) classifiers, which perform better by predicting based on the conditional probability of a label, given an input (Bird et al., 2009). The conditional probability is also calculated from the joint probability for the generative models. Error analysis is useful in refining the featureset (model inputs) as it provides the opportunity to know where the classifier excels and where it struggles.

1.7 Idioms

An idiom is a Multi-Word Expression (MWE) that has a different meaning from the constituent words that make it up (Quinn and Quinn, 1993; Drew and Holt, 1998). It may also be a word used in an abstract form instead of the literal sense. Not every MWE is an idiom, however. A compositional MWE gives away its meaning through the meaning of its composite words (Diab and Bhutada, 2009). Idioms are part of figures of speech, though some hold a different view, preferring to distinguish between the two (Grant and Bauer, 2004). Their usage is quite common in speech and written text (Lakoff and Johnson, 2008; Diab and Bhutada, 2009). They are culture-centric and may not always be universal. This can make it challenging for people from a different background to understand some idioms from other cultures. Idioms, sometimes, may not be well-defined, leading to difficulty in classification (Grant and Bauer, 2004; Alm-Arvius, 2003). A single word, at times, may be expressed as a metaphor (Lakoff and Johnson, 2008; Birke and Sarkar, 2006). This further complicates figure of speech (or idiom) identification (Quinn and Quinn, 1993). Since we recognise that idioms are a subset of figures of speech, we use figures of speech and idioms interchangeably, in this work. Examples of idioms are *"the nick of time"*, *"a laugh a minute"*,*"out of the blue"*, and *"dyed-in-the-wood"*, which are all metaphors. The examples mean *"just before the last moment"*, *"very funny"*, *"unexpectedly"*, and *"unchanging in a particular belief"*, respectively. Idioms pose challenges in various NLP tasks, including NLU, WSD, IR, conversational systems, and MT (Korkontzelos et al., 2013; Mao et al., 2018). Below are six examples of the difficulty the Google MT system experienced while translating sentences that have idioms from English to Swedish and then back again to English.

1. "but when we get to the end of the month, it's crunch time, " she says
   Translation ->
   "men när vi kommer till slutet av månaden är det dags för kris", säger hon
   Back-Translation->
   "but when we get to the end of the month, it's time for crisis," she says

2. 'You have come in the nick of time,' Alexandra told him
   Translation ->
   "Du har kommit i snäppet", sa Alexandra till honom
   Back-Translation->
   "You've been caught," Alexandra told him

3. I'm just a laugh a minute, Moses. You should keep me around and find out.
   Translation ->
   Jag är bara ett litet skratt, Moses. Du borde hålla mig runt och ta reda på det
   Back-Translation->
   I'm just a little laugh, Moses. You should keep me around and find out.

4. she arrived at lunch time, out of the blue to us
   Translation ->
   hon anlände vid lunchtid, direkt till oss
   Back-Translation->
   she arrived at lunchtime, directly to us

5. Stahl belongs to that dyed-in-the-wool amateur breed
   Translation ->
   Stahl tillhör den infärgade amatörrasen
   Back-Translation->
   Stahl belongs to the colored amateur breed

6. The business I've just bought is on the rocks
   Translation ->
   Verksamheten jag just har köpt är on the rocks
   Back-Translation->
   The business I just bought is on the rocks

In conversational systems, a user may appreciate a chatbot that identifies and generates an appropriate and better response based on the the idiom in a prompt than one that does not. For example, *"My wife kicked the bucket"* should have different responses from a conversational system, depending on the identification of the MWE as a literal usage or a specific idiom type, in this case, euphemism (a polite form of a hard expression). Correctly identifying the specific type of idiom instead of a general identification may elicit an empathetic response from the conversational system for the euphemism example. In addition, such classification has the potential benefit of automatic substitution of the idioms with the literal meaning for MT for the target language.

Idiom classification

Attempts at classifying idioms fall into different approaches like semantic, syntactic, and functional classification (Grant and Bauer, 2004; Cowie and Mackin, 1983). As depicted in Figure 1.3, classification of idioms can sometimes overlap (Grant and Bauer, 2004; Alm-Arvius, 2003). Classification of a case as euphemism also fulfills classification as metaphor. This is also the case with apostrophe. Therefore, two annotators with such different annotations may not imply they are wrong but that one is more specific. Metaphor uses a type of experience to outline something that is more abstract (Alm-Arvius, 2003; Lakoff and Johnson, 2008). It describes an entity by comparing it with another dissimilar thing in an implicit manner. Simile, on the other hand, compares in an explicit manner. Personification ascribes human attributes to inanimate things. Apostrophe denotes direct, vocative addresses to things which may not be factually present (Alm-Arvius, 2003). Contradictory combination of words or phrases is an Oxymoron. They are paradoxically meaningful and may appear hyperbolic (Alm-Arvius, 2003). Hyperbole is an overstatement and it has the effect of startling or amusing the hearer. Section 2.4 discusses about additional examples of idioms and the PIE-English idioms corpus (Adewumi et al., 2021). Figure 1.3 is a schematic representation of the relationships among some common idioms, based on the authors' perception of the description by Alm-Arvius (2003).

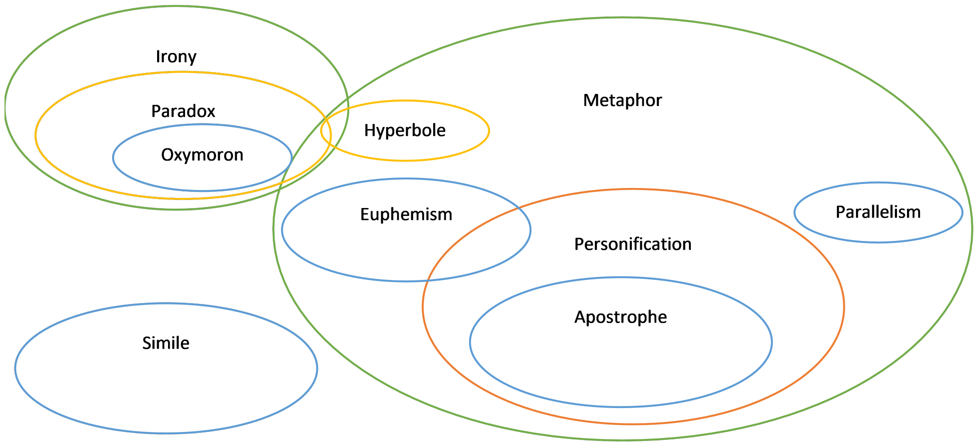

Figure 1.3: Relationship among some classes of idioms (Adewumi et al., 2021).

1.8 Scientific method

The scientific approach is based on evidence through experiments or empiricism for acquiring knowledge. It consists of an iterative sequence of principles that is applicable to all scientific endeavours. The basic, general components are shown in Figure 1.4. It starts off with careful observation, which requires rigorous skepticism through questions, then hypotheses formulation through induction (based on what has been observed), testing by experimentation, analysis of findings and, lastly, refinement of the hypotheses, as a result of the findings from the experiments (Newton, 1833). There are slightly different versions of the approach, especially as it concerns different scientific endeavours. According to Galilei (1954), the scientific approach also includes other components that are required even when all the stages identified in Figure 1.4 have been completed. These components are replication, external review, data recording, and sharing. The last one is essential for the first one (Fleck, 2012). The more specific process for this work is depicted in Figure 1.5.

The scientific method requires that the hypothesis is tested in controlled conditions whenever possible. Experimental control and reproducibility have the effect of reducing the misleading effect of circumstance and personal bias, to a certain degree, as (confirmation) bias can alter the interpretation of results (Javed et al., 2021; Snyder, 1984; Suresh and Guttag, 2021). The confirmation bias acts as some heuristic that leads someone to find things that reinforce their beliefs though another person may objectively observe otherwise (Snyder, 1984). We use models to simulate experiences. When such simulation of a model is assessed as similar to its actual counterpart in certain dimensions, it is considered successful (Colby et al., 1972).

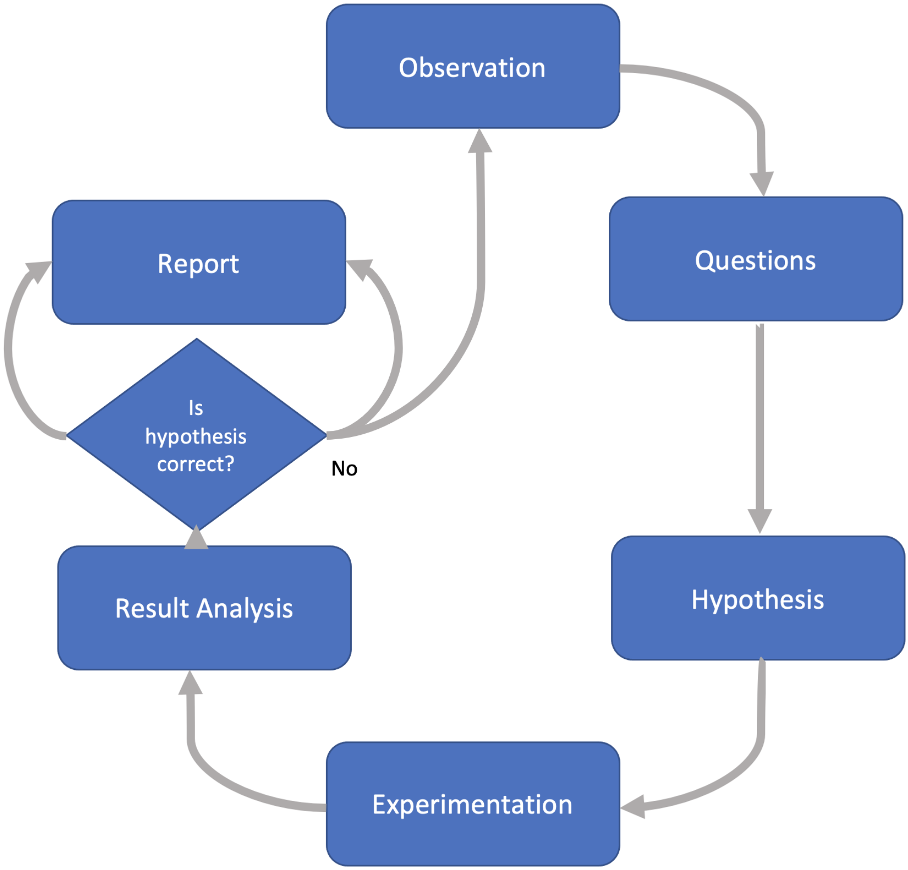

Figure 1.4: The general scientific approach

For a hypothesis to be considered scientific, it should be falsifiable (Popper, 2005; Adewumi et al., 2019). This means there should be an identification of a possible outcome of the experiment that conflicts with predictions from the experiment, based on the hypothesis. If this is not the case, then the hypothesis cannot be tested meaningfully. From the conjecture (or hypothesis), after an observation, we usually construct the null hypothesis and an alternative hypothesis (Du Prel et al., 2009). The null hypothesis assumes the relationship or the effect being examined is not really there, i.e., it is zero (Frick, 1995). It assumes sampling error is the reason for experiencing any difference in

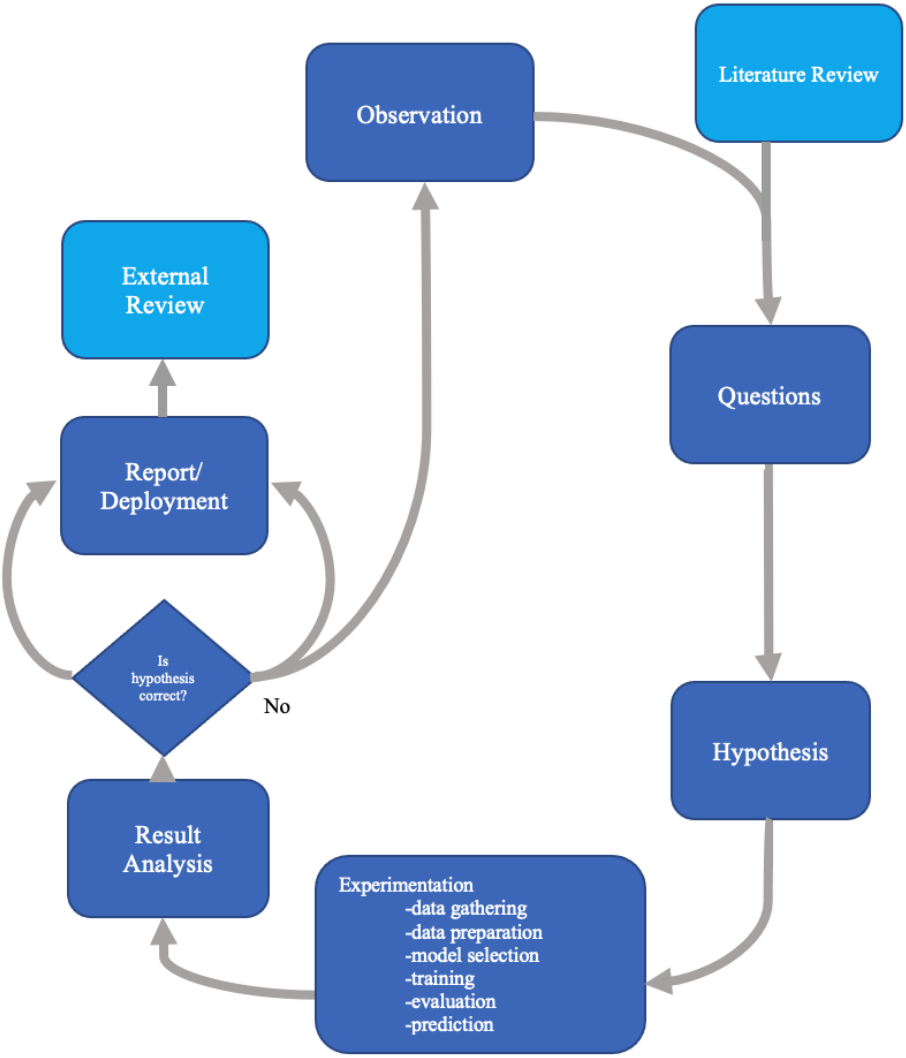

Figure 1.5: Methodology of this study

the data. However, the alternative hypothesis assumes there is truly a relationship or a nonzero effect or difference. Analysis of the data may be carried out using tools like regression, comparison of means using t-test, and analysis of variance. We may test the condition that if the null hypothesis is true, can one get an observed effect since we can not test if the null hypothesis itself is true. P-value is the probability of getting a result with observed effect if the results are due to chance or the null hypothesis is true. In other words, p-value = $P(data|nullhypothesis = true)$. Therefore, a high p-value means the result is easily due to chance and is statistically insignificant while a sufficiently low p-value, against the chosen alpha value, means the result is not easily due to chance and is statistically significant (Du Prel et al., 2009; Nickerson, 2000).

The alpha value sets a threshold for the types of errors that may occur. The type I error occurs when one detects an effect or relationship when actually there is none, resulting in false positives, while type II error occurs when no effect is detected though actually there is, resulting in false negatives. It must be noted that a low p-value does not mean we have proven a case. Rather, a low p-value implies the data or the null hypothesis is likely wrong because they are incompatible so we choose to trust our data and reject the null hypothesis. Noteworthy that there are several objections to hypothesis testing (Frick, 1996; Nickerson, 2000). Nickerson (2000) found that when there are no estimates of mean or the effect size, then null hypothesis testing is of no value. They further assert that null hypotheses testing have relatively little utility and are not part of the scientific approach. They, therefore, recommended data analysis that is based on Kullback-Leibler information instead of null hypothesis testing, though they pointed out that this is not perfect either.

Another statistical approach is to use confidence interval (CI) . This has the advantage of providing more information about the result than the p-value (Du Prel et al., 2009). It gives a range of the differences or the effect. In statistical tests, in order to draw valid conclusions, it is crucial to consider *"power"* and not filter out non-significant findings. It is also important to determine the *"power"* of an experiment or observation early on. It is dependent on the effect size and the size of the sample (Brysbaert and Stevens, 2018). It reflects the number of times the null hypothesis may be rejected or the ability to reject the null hypothesis. A large effect size with a relatively small sample size or a large sample size with a minimal effect size will result in good *"power"* (Brysbaert and Stevens, 2018).

The importance and difference between reliability, which is to rightly measure something, and validity, which is to measure the right thing, should be kept in mind. For results to be reliable, one should minimise errors that are due to survey measurement, which are errors captured with what is being measured and is common with latent measurements, such as sentiments, that have to be inferred. Latent measurement is different from manifest measurement that are measured directly, such as height or sales (Skrondal and Rabe-Hesketh, 2007).

Methodology

The specific methodology employed in this thesis involves an iterative set of scientific methods. As depicted in Figure 1.5, a systematic literature review is conducted to identify the state-of-the-art (SoTA) and gaps in current research. We acquire or create relevant datasets using benchmark datasets as references, as the need may be. Their data statements are documented as well. The seven stages of the machine learning life-cycle are followed as the datasets are used to train model architectures for predictions. The stages include data acquisition, data preparation, model selection, training, evaluation with hyperparameter tuning, prediction, and model deployment (Suresh and Guttag, 2021). We conduct human evaluation on the generated predictions of some of the conversational models. The results of such evaluation and comments from evaluators provide valuable feedback on challenges that may still exist within the system. Saygin and Cicekli (2002) show that when conducting tests (or evaluation), similar to the Turing test, knowledge of whether a machine is one of the respondents makes a difference in the evaluation by the judges. However, during the evaluation of PARRY, this information was not considered important (Colby et al., 1972). The knowledge works against the machines during evaluation by judges as shown by Saygin and Cicekli (2002).

Details of the implementation of the experiments to determine the status of the hypotheses of this work (Section 1.5) are provided in the various sections that follow this chapter. Experiments were run on a shared cluster running the Ubuntu operating system with multiple V100 GPUs, each having 32G memory. Preprocessing, such as removal of punctuation marks and lowering of cases, where appropriate, is applied to data before training. We perform multiple runs of each experiment and then report the average values. For tuning hyperparameters, grid search may be used. It is an exhaustive search that explores all possible combinations of the values supplied. The values may be computed in parallel to reduce the computational cost involved (Hackeling, 2017).

1.9 Performance metrics

We have to measure a system to ascertain the performance of such a system. There are a wide variety of metrics for NLP systems (Aggarwal and Zhai, 2012; Gehrmann et al., 2021; Reiter, 2010) but different metrics may be suitable for different systems, depending on the characteristics of the system. For example, IR systems may use F1, precision, and recall (Aggarwal and Zhai, 2012). We shall only mention a few of the possible NLP metrics here, some of which are used in this work. Human evaluation is the *gold standard* when it comes to the evaluation of conversational systems. It is, however, time-intensive and laborious. Consequently, automatic metrics serve as timely proxies for estimating performance though they may not correlate adequately with human evaluation (Gehrmann et al., 2021; Gangal et al., 2021; Jhamtani et al., 2021). Two methods of human evaluation may be conducted on open-domain conversational systems: observer and participant evaluation (Jurafsky and Martin, 2020). Observer evaluation involves reading and scoring a transcript of human-chatbot conversation while participant evaluation directly interacts with the chatbot in a dialogue (Jurafsky and Martin, 2020).

An open-domain conversational system may be evaluated for different qualities, such as humanness (or human-likeness), engagingness, fluency, making sense, interestingness, avoiding repetition, and more. The use of automatic metrics, such as the BLEU or ROUGE (Lin, 2004; Papineni et al., 2002), for evaluation of chatbots is sometimes viewed as inappropriate (Liu et al., 2016). This is because BLEU and similar metrics do poorly in measuring response generation, as they do not correlate well with human assessment, especially as they do not take lexical or syntactic variation into consideration (Reiter, 2010). Dependency-based evaluation metrics allow for such variation in evaluation. Another common metric for conversational systems is perplexity (Adiwardana et al., 2020). It measures how well a probability model predicts a sample and corresponds to the effective size of the vocabulary (Aggarwal and Zhai, 2012). Therefore, smaller values show that a model fits the data better. More is discussed about this in Section 4.4. Perplexity correlates with entropy (information gain). Entropy measures the amount of information in a random variable. It is the average uncertainty of a single random variable. The more we know about a variable, the lower the entropy, as we become less surprised by the outcome of a trial (Aggarwal and Zhai, 2012).

Evaluation of NLP systems may be achieved at two levels: intrinsic and extrinsic levels (Reiter, 2010). Unlike extrinsic metrics, intrinsic metrics do not capture the usefulness of a system in the real world but act as possible proxies (Reiter, 2010). Extrinsic evaluation methods focus on the usefulness of models with regards to downstream NLP tasks, such as Named Entity Recognition (NER) (Wang et al., 2019). The common metrics for extrinsic evaluation include accuracy, precision, recall, and the F1 score (Gatt and Krahmer, 2018). They are represented mathematically in Equations 1.1, 1.2, 1.3, and 1.4, respectively, using the concepts of true positive (TP), which is the number of items correctly classified as positive instances, true negative (TN), which is the number of items correctly classified as negative instances, false negative (FN), which is the number of items incorrectly classified as negative instances, and false positive (FP), which is the number of items incorrectly classified as positive instances. Precision tells us how often the system is correct when the system predicts the positive result. Recall tells us how often the system predicts correctly when it is actually the positive result. The F1 score is the harmonic mean of both the precision and recall (Aggarwal and Zhai, 2012; Powers, 2020). Accuracy can be misleading when used for search tasks, since a model that labels every irrelevant document in a retrieval system would be close to 100% (Bird et al., 2009). The visualisation metric receiver operating characteristics (ROC) - area under the curve (AUC) also depend on the concepts of true positives, true negatives, false positives, and false negatives. The confusion matrix presents a good visualisation of tagging errors by charting gold standard tags against actual tags generated by the tagger (Bird et al., 2009; Hackeling, 2017).

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{1.1}$$

$$\frac{TP}{TP+FP} \tag{1.2}$$

$$\frac{TP}{TP+FN} \tag{1.3}$$

$$\frac{2TP}{2TP+FP+FN} \tag{1.4}$$

1.10 Ethical consideration

From the viewpoint of deontological ethics, it is important to be objective in research (Javed et al., 2021; White, 2009). Deontological ethics is a philosophy that emphasis responsibility or duty over the ends achieved in decision-making (Alexander and Moore, 2007; Paquette et al., 2015). It has the advantage of accounting for moral intuitions than other viewpoints, like consequentialism, however, it has its disadvantages, such as the possibility of conflict of duties (Paquette et al., 2015). The Foundation and Academies (2017) identifies four guiding principles of research: reliability, honesty, respect, and accountability. This work adheres to those four principles, the good research practices that they prescribe, and the General Data Protection Regulation (GDPR). The GDPR is a regulation that protects natural persons with regards to the processing of their personal data and on the free movement of such data (Voigt and Von dem Bussche, 2017).

Ethical issues are of importance in open-domain conversational systems. Some of the issues that should be considered are privacy concerns arising from personally identifiable information (PII), toxic/offensive/hateful messages that may surface as a result of the training data and bias (be it gender, racial, or other forms of bias) (Jurafsky and Martin, 2020). The data used for pretraining the deep models or embeddings in this work are from online public sources that are known to contain all kinds of views and they suffer from the risks identified. Therefore, we note that there are risks with using the produced model checkpoints or embeddings, as they may show such biases or offensive language (Zhang et al., 2020).

1.11 Delimitation

This work is the intersection of multilingual NLP, idioms, and open-domain conversational systems. The thesis does not go into the details of the philosophy of language and linguistics, especially as described by Bach and Harnish (1979). It also does not discuss the details of conversational analysis (Sacks et al., 1978). We do not cover all possible combinations of hyperparameters for a given ANN and we cover only a few NLP downstream tasks. It is not practical to cover all possible hyperparameter combinations, as the combination increases faster than linearly with each additional hyperparameter factor. Also, this work does not experiment with all shallow neural networks for embeddings; it does not explore all deep models for conversational systems nor does it cover all NLP downstream tasks. Also, we acknowledge that figures of speech or idioms are so diverse that a detailed evaluation is out of the scope of this work. Finally, the discussion about open-domain conversational systems only prepares the ground for ongoing and future work. It highlights factors, which are important for ethical and robust open-domain conversational systems from the point of view of the philosophy of science (Adewumi et al., 2019).

1.12 Related work

Jhamtani et al. (2021) observed huge performance drop, with regards to figurative language, when they evaluated some deep models on two open-domain dialogue datasets: DailyDialog and PersonaChat (Li et al., 2017; Zhang et al., 2018b). Generative Pretrained Transformer (GPT)-2 was compared to four other models over the datasets and considerable drop in performance was observed in most. Their approach of transforming figurative language (including idioms) to their literal form before feeding the model may not adequately address the challenge since this implies the models still are incapable of understanding the figurative language and because some idioms have more than one literal form.

Zhang et al. (2020) pretrained the deep model, DialoGPT, on conversational data from Reddit conversations of 147M exchanges. The model, which comes in three different flavours, achieved performance close to that of humans in open-domain dialogues of single-turn conversations. DialoGPT is based on GPT-2 (Radford et al., 2019). Hu et al. (2018), Olabiyi and Mueller (2019), Adiwardana et al. (2020), and Roller et al. (2021) pretrained their models, Texar, DLGnet, Meena, and BlenderBot respectively, on dialogue datasets also. Some architectures are pretrained on large, semi-structured (or unstructured) text and adapted for conversational systems. These include T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). Xu et al. (2017) found that a deep LSTM-based model outperformed a standard IR baseline for response generation to customer requests for about sixty brands on social media but achieved similar performance as humans in handling emotional situations.

Different methods have been employed in past efforts for creating idioms corpora. Some of the labelled idioms datasets available only focus on two categories (or senses of expressions): the literal and general idioms classification (Li and Sporleder, 2009; Cook et al., 2007). Sporleder et al. (2010a) presented the IDIX corpus, which has 78 idioms in 5,836 sentence samples. They identify five categories for labelling the samples: literal, non-literal, both, meta-linguistic, undecided. To create the corpus, they pick selections in idiom dictionaries and use Google to know how frequent each idiom is. Then, they search the BNC online to determine examples of literal and non-literal. They went in favour of expressions that are frequent online, that are in the BNC and have idiomatic and literal meanings. Instead of manually curating the expressions, a perl script was used to automatically extract all occurrences of desired expressions from the BNC and erroneous extractions manually filtered out during annotation (Sporleder et al., 2010a). Meanwhile, Cook et al. (2007) selected 60 verb-noun construct (VNC) token expressions and extracted 100 sentences for each from the BNC. These were annotated using two native English speakers (Cook et al., 2007). Saxena and Paul (2020) introduced English Possible Idiomatic Expressions (EPIE) corpus, which has 25,206 samples of 717 idiom cases. Haagsma et al. (2020) generated potential idiomatic expressions (MAGPIE) and annotated the dataset using only two main classes (idiomatic or literal), through crowdsourcing. The samples of idioms are 2.5 times more frequent than the literals. It has 1,756 idiom types, an average of 32 samples per type, 126 types with only one instance and 372 cases with less than 6 instances.

Two approaches are common for idiom detection: type-based and tokens-in-context (or token-based) (Peng et al., 2015b; Cook et al., 2007; Li and Sporleder, 2009; Sporleder et al., 2010b). The type-based approach attempts to distinguish if an expression is an idiom, possibly through automatic compilation of an idiom list from a corpus (Sporleder et al., 2010a), while the token-based approach relies on context for disambiguation of idioms (Korkontzelos et al., 2013; Sporleder et al., 2010b). Non-contextual word embeddings (like word2vec) are used for identifying metaphors (Mao et al., 2018), which may then be used for additional downstream tasks, like MT. Such approaches are likely to underperform, however (Mao et al., 2018). Peng et al. (2015a) use word2vec to obtain vectors from text8 corpus with a vector dimension of 200. Their algorithm uses inner product of context word vectors with vector representing target expression. This is based on the assumption that literal vectors are distinguished from idiom vectors by the larger inner product they produce. The scatter matrices represent context distributions, which can be measured using Frobenius norm. Bizzoni et al. (2017a) employ word2vec and an ANN with 1 hidden layer for detecting metaphors. The corpus that the work is based on eliminated all adjective-noun (AN) phrases that require a longer context for their interpretation. Diab and Bhutada (2009) used support vector machine (SVM) to perform binary classification into literal and idiomatic expressions on a subset of the VNC-Token. In addition, Shutova et al. (2016) describe using textual and visual clues for metaphor identification.

In evaluating the performance of open-domain chatbots, it has been shown that automatic metrics, like the BLEU score, can be poor but they are still used in some cases (Lundell Vinkler and Yu, 2020). Conversation turns per session is another metric of interest (Zhou et al., 2020a). Perplexity is also widely used for intrinsic evaluation of language models and its theoretical minimum, which is its best value, is 1 (Adiwardana et al., 2020). Gangal et al. (2021) reiterated that previous work reveals the importance of having multiple valid responses as reference for meaningful and robust automated evaluations. Perhaps the best evaluation is done by humans though this can be subjective. Human judgment is seen as very important, since humans are usually the end-users of such systems (Zhang et al., 2020).

1.13 Thesis Outline

The remaining five chapters cover *data*, *vector space*, *open-domain conversational systems*, *learning deep abstractions*, and *conclusion and future work*. Chapter two is dedicated to *data*. We discuss in detail the datasets we created, their methodology, and their data statements. These include the Swedish analogy test set and the PIE-English idioms corpus. In addition, it discusses the AfriWOZ dataset, which are translations of the multi-domian MultiWOZ dataset. The chapter ends by describing the results of experiments on classifiers used for the PIE-English idioms corpus.

Chapter three, which discusses *vector space*, provides details of experiments on word vectors, contextual versus non-contextual representation, and evaluation of embeddings.

Chapter four discusses the differences between open-domain and task-based systems, deep models for open-domain chatbots, evaluation of conversational systems, and the ethics of building dialogue systems. It also discusses the new Credibility unanimous score (CUS) for calculating IAA. Chapter five, which is *learning deep abstractions*, highlights some commonalities in human languages, discusses the issue of pretraining, looks at the pros and cons of multi-lingual deep models, and the experimental results on cross-lingual transferability for the various languages tested. The final chapter concludes this work by reiterating important points, contributions, and possible future work. Figure 1.6 gives a depiction of the structure of this thesis.

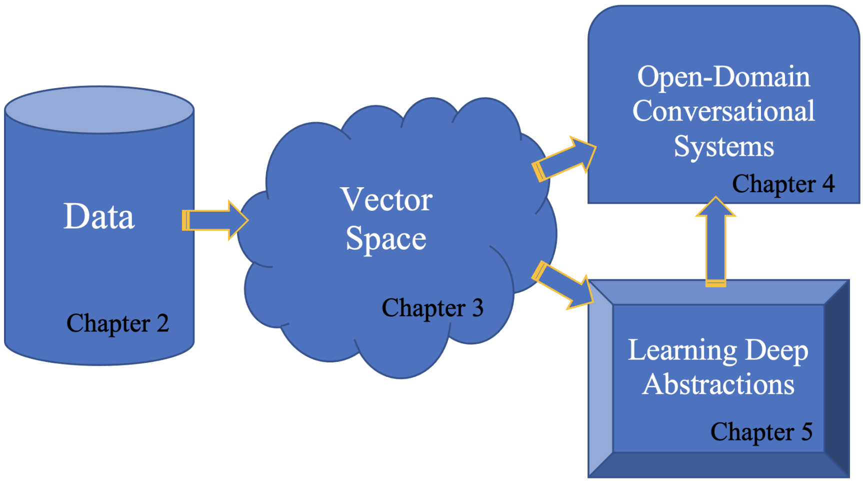

Figure 1.6: Schema of the structure of this thesis

Chapter 2

Data

"Data is the new oil."

(Metaphor)

Data is, perhaps, the most important ingredient in the ML life-cycle. In order to train ANN we need data. If data can be scarce, quality data can be more scarce, especially, quality labelled data (Crawford et al., 2015). This is more so for low-resource languages, such as Yorùbá, Igbo, Hausa, Wolof, and many more. Textual data can be in many formats and may be available in different media. The type of data and size that is available can determine the type of training and the architecture that such data may be used with. Large, unstructured or semi-structured textual data may be used in the pretraining of deep ANNs (Devlin et al., 2018a; Raffel et al., 2020). Typically, a task-specific dataset, which may be labelled, is first divided into 2 main categories: the development (dev) and test sets. The development set is then further divided into the training set and the final dev (or validation set). The final ratio of the split is based on a tradeoff. The test set should not be too small, as it may be unrepresentative of the training set, so it should be large enough to give statistical power. Meanwhile, we want the training set to be as much as possible for the model to learn from as many samples as possible (Jurafsky and Martin, 2020). The final ratio might be around 80:10:10, such that the dev set is representative of the test set. The dev set is used to perform error analysis after each epoch of training, which is very useful for refining the featureset (Bird et al., 2009).

A shuffling of the training and dev sets is important each time error analysis is repeated to avoid overfitting (Bird et al., 2009). The method of cross validation, where multiple evaluations are conducted on various test set splits from the same dataset and the results combined, has two advantages: it is useful for cases when the entire dataset is small and allows assessment of how widely performance varies across the different test sets used. With good similarity in the scores of the number of sets used, there's confidence in the accuracy of the score (Bird et al., 2009). Furthermore, the test set is recommended to have, at least, fifty instances of the infrequent label, if a corpus (or dataset) for a classification task has infrequent labels. Usually, it can be difficult for a model to generalise to other datasets when the training and test sets are very similar.

However, using a more stringent evaluation set, sometimes referred to as a challenge set (Gehrmann et al., 2021), the test set may be transformed or drawn from a different category of documents slightly less related to the training set. Some of the datasets available for training classifiers include the Internet Movie Database (IMDB) (Maas et al., 2011), CoNNL-2003 (Aggarwal and Zhai, 2012), and the Groningen Meaning Bank (GMB) (Bos et al., 2017) while examples of those available for training conversational systems include the BlendedSkillTalk (BST) (Smith et al., 2020), and MultiWOZ (Budzianowski et al., 2018).

In dataset creation, it is unlikely that one covers every possible scenario or instances with every possible attribute. The Wizard-of-Oz (WOZ) approach to data creation, where participants interact with a presumed automated system, which in reality is simulated by an unseen human participant, appears to be common (Byrne et al., 2019; Budzianowski et al., 2018; Jurafsky and Martin, 2020). It is an imperfect approach that may not capture the real limitations or constraints of the system being simulated but provides a useful step towards data acquisition (Jurafsky and Martin, 2020). In cases where there is a lack of diversity or imbalance in the dataset, it is better to take measures to increase the dataset so as to avoid a skewed dataset and evaluation. Datasets may be annotated for several properties. For example, a speech dataset may be annotated for phonetic while a sentiment dataset may be annotated for positive and negative sentiments (Bird et al., 2009).

The rest of this chapter is organised as follows: Section 2.1 discusses how data acquisition may be carried out. Section 2.2 discusses the issues around IAA. Section 2.2 gives details about the Swedish analogy test set, one of the contributions of this thesis. Section 2.4 gives details of the PIE-English idioms corpus, another contribution of this thesis. Section 2.5 discusses details of the six datasets combined as AfriWOZ, which is another contribution of this thesis. Section 2.6 discusses the importance of data statements. Section 2.7 shows results from experiments conducted on idiom classification.

2.1 Methodology of data acquisition

Building a dataset requires time and careful preparation. Depending on the type of data and the task it is meant for, different stages may be involved in the dataset acquisition process. The process may involve (automatic or manual) annotation and post-editing. (Bird et al., 2009) Figure 2.1 shows a depiction of the possible stages of data acquisition. The stages in the figure are by no means exhaustive and may be refined as the application warrants. The three common approaches for data acquisition are data discovery, data augmentation, and data generation (Roh et al., 2019). Data discovery is applicable when there's data available on the web or other sources from which one may search and acquire the dataset. When data is acquired through data discovery, one might augment the existing data in order to complement it. For example, the subsequent MultiWOZ datasets (Eric et al., 2019) that built on the original by Budzianowski et al. (2018) are examples of this approach. The third approach of data acquisition involves the manual (through crowdsourcing or otherwise) or synthetic means of generating data when it is not available (Roh et al., 2019). The first MultiWOZ dataset by Budzianowski et al. (2018) is an example of this.

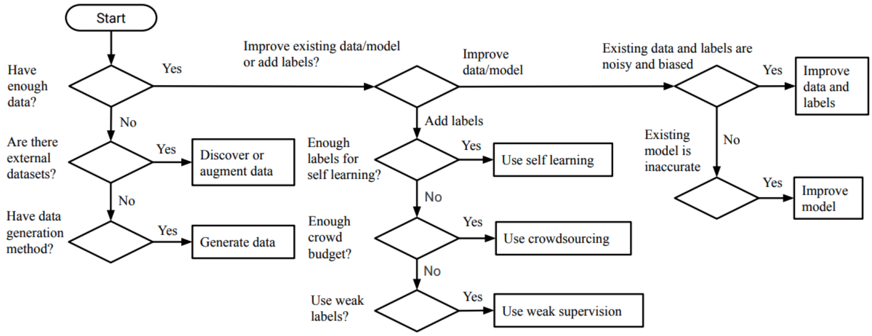

Figure 2.1: Stages of data acquisition (Roh et al., 2019).

Uncertainty with some samples during annotation may require adjudication, when labelling or augmenting data. Having a documentation to accompany the dataset, training of the workers involved in the dataset acquisition and procedure for the data acquisition will impact on the quality of the dataset. Versioning is an important part of the process of data acquisition (Bird et al., 2009). If the data acquisition involved annotation, best practice requires that the IAA be reported. This IAA is usually perceived as the upper bound on the expected performance of ML models that are trained on the corpus (Bird et al., 2009; Clark et al., 2012).

2.2 Inter-Annotator Agreement (IAA)

As humans, we have subjective views, which may influence our decisions, even when annotating or labelling data, though there may be an annotation guide (or document). This is why it is good practice to have more than one person labelling such data and to calculate their IAA. This agreement score is a requirement after the process of annotation (Peng et al., 2015b). In general, one might expect that with more annotators the consistency of annotation increases - and this is sometimes the case. However, if there are experts and non-experts involved, problems may arise. Another situation that may arise during annotation is a tie (deadlock), i.e., when an item is labelled differently by the same amount of annotators (Bird et al., 2009). Using odd number of annotators usually resolves the deadlock problem. A typical measure to improve annotation consistency is to provide annotators with an annotation guide. The annotation guide, which will detail the rules for the task of annotation should be simple enough for many to follow and be objective. This will help to reduce instances of confabulation among annotators, when people make up false reasons unintentionally for doing something or making certain choices, and increase the chances of high IAA. The task of annotation is either too difficult or poorly defined (possibly from the guide) if the annotators are not able to achieve good enough agreement on the correct annotations (Clark et al., 2012).

A simple way of measuring IAA agreement among annotators is to measure their observed percentage agreement over the data samples. However, this method may be an inaccurate reflection of the actual difficulty or upper bound on the task, as some agreement may be due to chance (Clark et al., 2012). Cohen's kappa and Fleiss kappa (k) are widely-used methods for calculating IAA. However, both have limitations, scope, and interpretation difficulties (Clark et al., 2012; Gwet, 2014; Landis and Koch, 1977). Fleiss (k) scores are lower when the number of classes or categories under consideration increases (Sim and Wright, 2005). A contribution of this thesis is the introduction of CUS for measuring IAA in open-domain conversational transcripts and this is discussed further in Section 2.2

Credibility unanimous score (CUS)

Raw percentages of observed agreement on a sample of annotated entities for measuring IAA has been shown to be weak since some agreements may be due to chance (Clark et al., 2012). Fleiss Kappa (k), another common IAA metric, has been shown to be restrictive in its interpretation, depending on the number of categories (Landis and Koch, 1977), as Kappa is lower when the categories are more (Sim and Wright, 2005). CUS is more intuitive, easier to calculate (as it's quite similar to raw percentages) and seemingly less sensitive to changes in the number of categories being evaluated, compared to Flies Kappa (k). The assumption behind CUS is that if homogeneous samples that are introduced can be used for establishing the credibility of the annotators for evaluating the dialogue transcript, then they may be used for establishing their agreement. This agreement is based on unanimous votes across the homogeneous samples. The homogeneous samples may be viewed as a significant subset of the full transcript, especially when it fulfils the central limit theorem by having a minimum of 30 samples. The probability of obtaining high CUS rises when the benchmark score for annotator credibility is raised. For example, if the benchmark scores for accepting annotators' work in two different jobs are 51% and 71%, then the probability of getting a higher CUS is higher in the latter. This gives CUS an advantage over using raw percentages over the actual samples, due to the weakness identified earlier.

2.3 Swedish analogy test set

Following the format of the original English analogy test set by Mikolov et al. (2013b), this thesis introduces the Swedish analogy test set (Adewumi et al., 2020c), with two main categories and their corresponding sub-categories: the semantic and syntactic sections. The task is to predict, per line, the forth item based on the third, given the similarity between the first and the second, as given in Figure 2.2. Many examples in the Swedish version are drawn from the English version. New entries were also added. The test set was constructed with the help of tools dedicated to Swedish dictionary/translation[1] and was proof-read for corrections by two Swedish native/L1 speakers (with an inter-annotator agreement score of 98.93%). Not all the words in the English version could be easily translated to Swedish, as similarly observed by Venekoski and Vankka (2017), while working on a smaller Finnish version. The English version has over 1,500 more syntactic samples than the semantic samples, however, the Swedish version is balanced across the two major sections and has more total samples, as shown in Table 2.1.

Table 2.1: The Swedish analogy test set statistics (Adewumi et al., 2020c)

Semantic	**Syntactic**
capital-common-countries (342)	gram2-opposite (2,652)
capital-world (7,832)	gram3-comparative (2,162)
currency (42)	gram4-superlative (1,980)
city-in-state (1,892)	gram6-nationality-adjective (12)
family (272)	gram7-past-tense (1,891)
	gram8-plural (1,560)

It has a total of 20,637 samples, made up of 10,380 semantic and 10,257 syntactic samples. The *capital-world* sub-category has the largest proportion of samples in the semantic subsection while the *gram6-nationality-adjective* sub-category has the least number of samples. Overall, there are 5 semantic subsections and 6 syntactic subsections. Table 2.2 provides some examples from some sub-categories of the dataset.

2.4 PIE-English idioms corpus

Figures of speech, which idioms are part of, become part of a language when members of the community repeatedly use it. The principles of idioms are similar across many languages but actual examples are not identical across languages (Alm-Arvius, 2003). The PIE-English idioms corpus has about 1,200 cases of idioms (with their meanings) (e.g. carry the day, add insult to injury, etc), 10 classes (or senses/categories, including literal), and over 20,100 samples from the British National Corpus (BNC)[2] , with 96.9%, and about 3.1% from UK-based web pages UKWaC (Ferraresi et al., 2008). The BNC has 100M words while the UKWaC has 2B words. This is possibly the first idioms corpus with classes of idioms beyond the typical literal and general idioms classification. Tables 2.3 and 2.4 show the distribution of the classes in the dataset and the annotation agreement, respectively. The total IAA score is 88.89%. Adjudication for the remaining 11.11% cases for the corpus was to accept the classification based on Alm-Arvius (2003). Table 2.5 shows some examples of sentences containing idioms in the corpus.

[1]https://bab.la & https://en.wiktionary.org/wiki/
[2]english-corpora.org/bnc

Table 2.2: Samples from some subsections in the Swedish analogy test set (Adewumi et al., 2020c). The task is to predict, per line, the forth item based on the third, given the similarity between the first & second.

:capital-common-countries			
Nassau	Bahamas	Havanna	Kuba
Nassau	Bahamas	Berlin	Tyskland
Nassau	Bahamas	Aten	Grekland
Nassau	Bahamas	Jakarta	Indonesien
Nassau	Bahamas	Jerusalem	Israel
Nassau	Bahamas	Rom	Italien
Nassau	Bahamas	Tokyo	Japan
Nassau	Bahamas	Nairobi	Kenya
:family			
pojke	flicka	bror	syster
pojke	flicka	far	mor
pojke	flicka	han	hon
pojke	flicka	hans	hennes
pojke	flicka	kung	drottning
pojke	flicka	farfar	farmor
pojke	flicka	man	kvinna
pojke	flicka	son	dotter
:gram2-opposite			
medveten	omedveten	lycklig	olycklig
medveten	omedveten	artig	oartig
medveten	omedveten	härlig	förfärlig
medveten	omedveten	bekväm	obekväm
medveten	omedveten	konsekvent	inkonsekvent
medveten	omedveten	effektiv	ineffektiv
medveten	omedveten	moralisk	omoralisk
medveten	omedveten	känd	okänd
:gram3-comparative			
dålig	sämre	stor	större
dålig	sämre	billig	billigare
dålig	sämre	kylig	kyligare
dålig	sämre	lätt	lättare
dålig	sämre	snabb	snabbare
dålig	sämre	bra	bättre

The idioms were selected alphabetically from the dictionary by Easy Pace Learning[3]

[3] easypacelearning.com

Table 2.3: Distribution of samples of idioms/literals in the corpus (Adewumi et al., 2021).

Classes	% of Samples	Samples
Euphemism	11.82	2,384
Literal	5.65	1,140
Metaphor	72.7	14,666
Personification	2.22	448
Simile	6.11	1,232
Parallelism	0.32	64
Paradox	0.56	112
Hyperbole	0.24	48
Oxymoron	0.24	48
Irony	0.16	32
Overall	**100**	**20,174**

Table 2.4: Annotation of classes of idioms in the corpus (Adewumi et al., 2021).

Classes	Annotation 1	%	Annotation 2	%
Euphemism	148	12.36	75	6.27
Metaphor	921	76.94	877	73.27
Personification	28	2.34	66	5.51
Simile	82	6.85	66	5.51
Parallelism	3	0.25	9	0.75
Paradox	6	0.5	19	1.59
Hyperbole	3	0.25	57	4.76
Oxymoron	4	0.33	9	0.75
Irony	2	0.17	19	1.59
Overall	**1197**	**100**	**1197**	**100**

and proverbs were excluded, since they are not the subject of the corpus. Samples of sentences were then extracted from the BNC and UKWaC, based on the first to appear in each corpora. Four contributors, who are second/L2 (but dominant) speakers of English, extracted sample sentences of idioms and literals (where applicable) from the BNC, based on identified idioms in the dictionary. The corpus was reviewed by a near-native speaker, as a form of quality control. This approach avoided common problems noticeable with crowd-sourcing methods or automatic extraction (Haagsma et al., 2020; Roh et al., 2019; Saxena and Paul, 2020). There are 2 sentences, at most, for each sample, though the majority of them contain only 1 sentence. Using one or two sentences minimises the possibility of having several different classes in one sample, which will make it difficult for annotation or classifiers to learn. The design involved having, for each idiom case, 15 samples and 21 for cases that have literal usage also, where 6 samples are literal for the cases that have literal usage. Six was chosen as the number of literal samples because the BNC and UKWaC sometimes have fewer or more literal samples, depending on the

Table 2.5: Samples from the PIE-English idioms corpus (Adewumi et al., 2021).

No	Samples	Class
1	Carry the day	Metaphor
2	Does the will of the Kuwaiti parliament transcend the will of the Emir and does parliament carry the day?	Metaphor
3	The anti Hunt campaigners believe they have enough to carry the day tomorrow	Metaphor
4	The pack particularly that controls the ball and makes fewer mistakes could carry the day	Metaphor
5	Time flies	Personification
6	Eighty-four!' she giggled.' How time flies	Personification
7	Think how time flies in periods of intense, purposeful activity	Personification
8	How time flies! We were at our stewardess's mercy	Personification
9	As clear as a bell	Simile
10	It sounds as clear as a bell	Simile
11	What you get is a sound as clear as a bell	Simile
12	It will make it as clear as a bell	Simile
13	Go belly up	Euphemism
14	If several clubs do go belly up, as Adam Pearson predicts.	Euphemism
15	That Blogger could go belly up in the near future	Euphemism
16	The laptop went belly up	Euphemism
17	The back of beyond	Hyperbole
18	There'd be no one about at all in the back of beyond.	Hyperbole
19	"Why couldn't you just stay in the back of beyond?" she said.	Hyperbole

case.

The BNC is a common choice for text extraction. It is, however, relatively small, hence we relied also on UKWaC for further extraction when search results were less than the requirements. Hence, there are 22 samples for each case that has literal usage, in addition to the original idiom while there are 16 for cases without literal usage. Metaphors, as expected, are the dominant cases in the PIE-English idioms corpus, which seems inevitable because metaphors are the most common figures of speech (Alm-Arvius, 2003; Bizzoni et al., 2017b; Grant and Bauer, 2004; Jhamtani et al., 2021). Part-of-speech tags are included for tokens in the corpus and this was performed by using the NLTK (Bird et al., 2009). The corpus may also be extended by researchers to meet specific needs. Table 2.6 compares the PIE-English idioms corpus with some other publicly available idioms datasets. The PIE-English idioms corpus has the largest number of classes, differentiating the many types of figurative speech that exist. It is also the third largest corpus, in terms of samples, and the second largest, in terms of cases.

*Table 2.6: Some datasets compared (*NA: not available) (Adewumi et al., 2021).*

Dataset	Cases	Classes	Samples
PIE-English (ours)	1,197	10	20,174
IDIX	78	NA*	5,836
Li & Sporleder	17	2	3,964
MAGPIE	1,756	2	56,192
EPIE	717	NA*	25,206

2.5 MultiWOZ to AfriWOZ

The MultiWOZ dataset has several versions, with each new one bringing improvements (Budzianowski et al., 2018; Eric et al., 2020). It is a fairly large, human-human, multi-domain, and multi-task benchmark conversational dataset. It has more than 10,000 dialogues distributed between multi-domain and single-domain dialogues. Domains covered include hospital, restaurant, police, attraction, hotel, taxi, train, and booking. AfriWOZ is a collection of conversational datasets in some African languages, based on translation of the Englsih MultiWOZ dataset. This data acquisition approach for AfriWOZ is needed because of the scarcity or non-existent conversational data for many African languages. The MultiWOZ seems better suited, as the source data, instead of alternatives like Reddit[4] because of the high probability of toxic content (Henderson et al., 2018; Roller et al., 2021). Solaiman and Dennison (2021) advocated for the careful curation of datasets as a safe approach to the adjustment of a model's behaviour to address the challenge of toxic comments. Such curation approach was used for the AfriWOZ. The first 1,000 turns from the training set and the first 250 turns each from the validation and test sets were translated from MultiWOZ to the 6 target languages: Swahili, Wolof, Hausa, Nigerian Pidgin English, Kinyarwanda & Yorùbá. Only 200 turns from the MultiWOZ training set were added to make up the 1,000 turns for the Yorùbá data because it has a small collection of conversational data online, which are a mix of short dialogues in different scenarios including the market, home and school. The two online sources[5] are used for Yorùbá because of the local entities in them.

Translation quality and challenges

The translators were recruited from Slack[6] and they are native/L1 speakers of the target languages and second/L2 (but dominant) speakers of English. Human translation was employed for all the languages except Hausa, which used Google MT. Review of all translations is then conducted for quality control (QC). The use of native speakers mitigated the risk of translating English conversations into unnatural conversations in the target languages. The two main human translation challenges encountered are how to handle English entities and how to reframe English conversations for cultural relevance in the

[4]reddit.com/
[5]YorubaYeMi-textbook.pdf & theyorubablog.com
[6]slack.com/

target languages. The entities in the data were retained since this may facilitate MT task. The cultural background of the native speakers made it relatively simple to frame the English conversations into seemingly natural conversations in the target languages.

2.6 Importance of data statements

Bender and Friedman (2018) advocates for data statements to be part of NLP systems by including them in papers that present new datasets or report work with datasets. A data statement (or card) is a structured set of statements describing the characteristics of a dataset, just as a model card is a structured set of statements describing the characteristics of a model. Model cards are discussed in Section 4.3.8. Data statements may be more important than model cards because ML models are, probably, useless without data. Failure to include data statements has possible consequences. Some of these consequences are poor generalisability of results, harmful predictions, and failure of NLP systems for certain groups. The failure can result from lack of representation or bias against such groups in the training data. Bias here refers to unwanted, systematic, and unfair discrimination (Adewumi et al., 2019; Bender and Friedman, 2018). These may be pre-existing biases in the society or technical biases (Bender and Friedman, 2018).

It is beneficial to have a short version and a long, detailed version, which may be linked from the short version (Bender and Friedman, 2018). The long version may contain details about a) curation rationale, b) language variety, c) demographics (including age, gender, race, etc), d) data characteristics, e) data quality, and other possible details that may be relevant. The short version of the data statement may be included in any use of the data and can be a summary of the details in the long version (Bender and Friedman, 2018). The short versions of the Swedish analogy test set, the PIE-English idioms corpus, and the AfriWOZ are given below. The long versions can be found in the appendices.

> ***Short data statement for the Swedish analogy test set.***
> This is the Swedish analogy test set for evaluating Swedish word embeddings.
> The licence for using this dataset comes under CC-BY 4.0.
> Total samples: 20,637
> Semantic samples: 10,380 (5 sections- capital-common-countries (342), capital-world (7.832), currency (42), city-in-state (1,892), family (272))
> Syntactic samples: 10,257 (6 sections - gram2-opposite (2,652), gram3-comparative (2,162), gram4-superlative (1,980), gram6-nationality-adjective (12), gram7-past-tense (1,891), gram8-plural (1,560))
> The long version of this data statement is in Appendix A.

> ***Short data statement for the PIE-English idioms corpus.***
> This is the Potential Idiomatic Expression (PIE)-English idioms corpus for training and evaluating models in idiom identification.
> The licence for using this dataset comes under CC-BY 4.0.
> Total samples: 20,174

There are 1,197 total cases of idioms and 10 classes.
Total samples of euphemism (2,384), literal (1,140), metaphor (14,666), personification (448), simile (1,232), parallelism (64), paradox (112), hyperbole (48), oxymoron (48), and irony (32).
The long version of this data statement is in Appendix B.

Short data statement for the AfriWOZ dataset.
This is the AfriWOZ dataset for training and evaluating open-domain dialogue models.
The licence for using this dataset comes under CC-BY 4.0.
Total natural languages: 6 (Swahili, Wolof, Hausa, Nigerian Pidgin English, Kinyarwanda & Yorùbá)
Total turns in the training set per language: 1,000
Total turns in the validation set per language: 250
Total turns in the test set per language: 250
Domains covered in the data include hotel, restaurant, taxi and booking.
The long version of this data statement is in Appendix C.

2.7 Experiments & Evaluation: Idioms classification

The PIE-English idioms corpus was split in the ratio 80:10:10 and trained on the BERT (Devlin et al., 2018a) and T5 (Raffel et al., 2020) pretrained models from the HuggingFace hub (Wolf et al., 2020). The base version of both models are used. The pre-processing involved lowering all cases and removal of all *html* tags, though none was found since the data was extracted manually and verified. Special characters and numbers were removed also. Shuffling of the training set is carried out before training. Batch sizes of 64 and 16 were used for BERT and T5, respectively. The total training epochs for both was 6. All experiments were performed on a shared cluster with 8 Tesla V100 GPUs, though only one GPU was used in training the models. Ubuntu 18 is the OS version of the cluster. From the results in Table 2.7, we observe that the T5 model outperforms the BERT model. It appears that the dataset is not overly challenging and this may be due to the choice of keeping the length of each sample at a maximum of 2 sentences. The p-value ($p < 0.0001$) of the two-sample t-test for the difference of two means (of the macro F1) is smaller than alpha (0.05), hence the results are statistically significant.

Table 2.7: Average accuracy & F1 results (sd - standard deviation)

Model	**Accuracy**		**weighted F1**		**macro F1**	
	dev (sd)	test (sd)	dev (sd)	test (sd)	dev (sd)	test (sd)
BERT	0.96 (0)	0.96 (0)	0.96 (0)	0.96 (0)	0.75 (0.04)	0.73 (0.01)
T5	0.99 (0)	0.98 (0)	0.98 (0)	0.98 (0)	0.97 (0)	0.98 (0)

Error analysis

Figure 2.2 shows the confusion matrix of the predictions against the true labels for the test set, using the T5 model. We observe that the model performs quite well even for classes that have few samples in the training set, such as *irony* and *hyperbole*. It struggles mostly in correctly classifying the *literals*, as it misclassified 9.3% of them as *metaphor*, possibly because it is the largest class in the dataset.

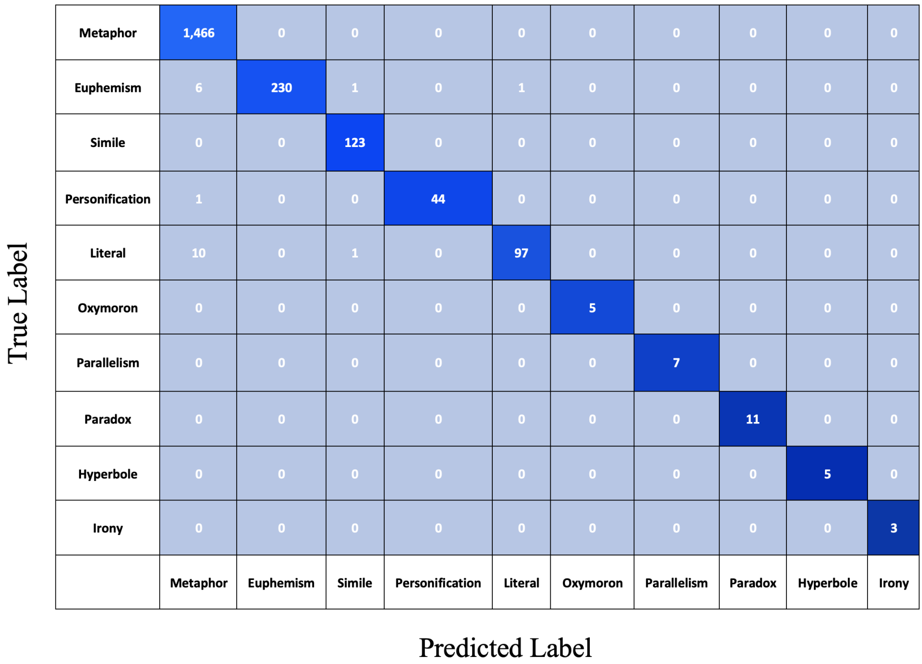

Figure 2.2: Confusion matrix for T5 model on the PIE-English test set.

Chapter 3

Vector Space

*"**The literature voices** different approaches to vector representation."*

(Personification)

Generally, a vector space model (VSM) represents each document, word or entity as a point (or vector) in a common space such that points that are close together are semantically similar. The converse is also true that points that are distant from one another are semantically distant (Manning et al., 2010; Turney and Pantel, 2010). The training corpus is divided into units, such as words or sentences, each of which is described by d-dimensional real-valued feature vector (Indurkhya and Damerau, 2010).

In this chapter, after discussing some background about VSM, Section 3.2 presents the curse of dimensionality. Thereafter, results from experiments using shallow neural networks in four experimental setups are presented in Section 3.3. Contextual versus non-contextual representation will then follow in Section 3.4 and the chapter will end with some more experiments on NER task for African languages.

3.1 Background

The VSM derives from the distributional hypothesis. The hypothesis describes how words that occur in a similar context tend to have similar or related meaning. It entails segmenting the words and ascertaining their similarity grouping (Harris, 1954; Firth, 1957). Hence, in a word-context matrix, words that have similar row vectors tend to have similar or related meaning (Turney and Pantel, 2010). VSM, based on linear algebra, underlie IR and treatment of word semantics, which is a search through a common space of states that represent hypotheses about an input (Jurafsky and Martin, 2020).

In Information Retrieval, the similarity of a set of documents and a query or another document determines the order of the result that is returned. These documents are sorted in order of increasing distance to the query (Salton et al., 1975). The maximum similarity is achieved when the angle between them is zero. The VSM relies on frequencies in the corpus for identifying semantic information. This practicality is based on the bag of words hypothesis (Salton et al., 1975). The hypothesis informs us that the relevance of a document to a query is indicated by the frequencies of words in that document. In a term-document matrix, when a document and the query have similar column vectors, there's the tendency they have similar meaning (Turney and Pantel, 2010). For term–document matrices, the term frequency-inverse document frequency (tf-idf) weighting functions formalise the idea that a surprising element has higher information content than an expected one (Shannon, 1948). When the corresponding term of an element is frequent in a document but scarce in other documents in the corpus, the element gets a high weight, as both the tf and idf will be high. TF-IDF weighting gives improvement over raw frequency. It's important to consider lengths of documents in IR to mitigate the bias which favours longer documents by performing length normalisation (Turney and Pantel, 2010). Performance in IR systems is usually measured by precision and recall (Manning et al., 2010). Apache Lucene[1] is an example of an open-source indexing and search software based on term-document matrix and provides additional features like spell-checking and analysis/tokenisation capabilities, which is used by Wikipedia and CNET (Turney and Pantel, 2010).

Prior to generating term-document or word-document matrix, application of some linguistic processing to the text is usually beneficial. Tokenisation is the first step, such that entities, words or subwords are extracted from the raw text (Harris, 1954), based on some algorithm. Normalisation may then follow. This process converts cases in one form to another (case folding), typically to lower case, and stems inflected words to their root form. In addition, it converts superficially different characters or entities to the same thing. For example, normalisation may involve replacing ö in öl with o, for the Swedish language, and bá in bábá with a, for the Yorùbá language. It is obvious that normalisation can distort original languages and may cause problems since case does have semantic significance in NER. The system finds it relatively easier to recognise similarities with normalisation, so recall increases while precision falls because of the error of variations. The final step may involve (automatically or manually) annotating entities in the text with additional information, such as parts of speech (Turney and Pantel, 2010).

The tokenisation step may appear simple for English text but an adequate tokeniser should also handle punctuation, hyphenation (such as state-of-the-art) and MWE (Manning et al., 2010). There are languages, such as Chinese, whose words are not separated by spaces. Hence, tokenisers specifically designed for English will not be adequate for such. In the tokenisation step, removal of "stop" words, which are frequently-occurring but relatively non-informative words, can be very good. Examples of "stop" words are 'the', 'of', and 'in'. The natural language toolkit (NLTK) by Bird et al. (2009) provides a list of "stop" words for English and some other languages. Obtaining highly accurate tokenisation is currently challenging for many human languages, as native speakers sometimes do not agree with the automatic segmentation produced (Turney and Pantel, 2010). Unlike normalisation, annotation adds additional information to entities in

[1] lucene.apache.org/

the data, hence, it may be viewed as the inverse of normalisation. It, therefore, has the reverse effects on precision and recall, and can provide better search results for a given query. This is useful for tokens with identical characters but which have different meaning (Turney and Pantel, 2010).

A very common way of ascertaining the similarity of two or more entities in VSM is through the cosine of the angle between them. It is the inner product of the vectors (say, $\mathbf{x}$ and $\mathbf{y}$) after normalisation to unit length, thereby making the length of the vectors irrelevant (Turney and Pantel, 2010). This is depicted in Equation 3.1.

$$cos(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{n}(x_i.y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2.\sum_{i=1}^{n}y_i^2}} = \frac{\mathbf{x}.\mathbf{y}}{\sqrt{\mathbf{x}.\mathbf{x}}.\sqrt{\mathbf{y}.\mathbf{y}}} = \frac{\mathbf{x}}{||\mathbf{x}||}.\frac{\mathbf{y}}{||\mathbf{y}||} \tag{3.1}$$

Its lower bound is -1, suggesting the vectors point in opposite directions in VS and its upper bound is +1, suggesting they point in the same direction. The cosine value is zero when the vectors are orthogonal. This measure of distance between vectors becomes a measure of similarity by subtraction or inversion, as given in Equations 3.2 and 3.3, respectively (Turney and Pantel, 2010). Although some classification and clustering algorithms can use cosine as a metric of similarity (Dasarathy, 1991; Jain et al., 1999), many ML algorithms work directly with the vectors in VSM (Turney and Pantel, 2010). A different approach to measuring similarity is by using information theoretic measure, like cross entropy, after a document is represented with a probability distribution over words (i.e. unigram language models) (Aggarwal and Zhai, 2012).

$$sim(\boldsymbol{x},\boldsymbol{y}) = 1 - dist(\mathbf{x},\mathbf{y}) \tag{3.2}$$

$$sim(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{dist(\mathbf{x},\mathbf{y})} \tag{3.3}$$

3.2 The curse of dimensionality

One of the early approaches of word representation was a bag-of-words (BoW), which accounts for the frequency of each term but is indifferent to the word order in a document, though it's simple (Aggarwal and Zhai, 2012; Mikolov et al., 2013b). Table 3.1 gives an example of this representation for the example sentence *'pat let the cat out of the bag'*. This method suffers from the large amount of components in the vector representation, thereby making it computationally relatively expensive. The representation retains document content and can be analysed with mathematical and ML techniques. However, the dimensionality of representation is usually very high, as each dimension corresponds to one term (Aggarwal and Zhai, 2012).

This large number of dimensions creates a problem for the task of analysis of concepts in documents. Typically, a low-dimensional space is preferred, where each dimension corresponds to one concept or feature. The ML technique of dimension reduction can be used to find the semantic space that reveals the preserved important properties of the corpus more clearly. It begins with a representation of the entities (usually using a BoW) and then finds a lower dimensional representation, which is considered faithful to the original representation. This feature transformation makes the features a linear combination of the features in the original data and removes noisy dimensions (such as synonymy and polysemy), which hamper similarity-based applications. Variances along the dimensions removed are small and the relative behaviour of the data points is minimally affected by removing them (Aggarwal and Zhai, 2012). The feature vectors represent different aspects of a word and the number of features is smaller compared to the vocabulary size (Bengio et al., 2003). Latent Semantic Indexing (LSI) is based on this feature transformation principle (Aggarwal and Zhai, 2012). Other useful applications based on the principle are Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). The standard matrix factorisation technique used by the earlier examples is different from the probabilistic framework for dimensionality reduction used by, say, Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic indexing (PLSI).

Table 3.1: Example of bag-of-words (BoW)

Term:	pat	let	the	cat	out	of	bag
Frequency:	1	1	2	1	1	1	1

The use of low-dimensional, distributed vectors (or embeddings) give more efficient representations (Mikolov et al., 2013b) compared to one-hot encoding or BoW, which represents each unique word as a single dimension. Tables 3.2 and 3.3 use the same example sentence provided earlier to show how the terms may be represented with one-hot encoding and low-dimensional representation, respectively. The one-hot encoding suffers from some of the issues BoW suffers from. These are data sparsity, poor semantic generalisation, low accuracy, and overfitting. Distributed representations derive from the distributional hypothesis, though the two words (distributed and distributional) are usually misunderstood or used interchangeably (Turian et al., 2010). Distributional word representation is the more general term, which is based on a co-occurence matrix F of size WxD, where W is the vocabulary size and D is the total dimension with some context. The choice of dimensionality being as large as the vocabuary, W, can be too large to use as features in a supervised model. However, mapping the initial matrix to a smaller one through a function such that the dimensionality of the new matrix is $d \ll D$ is usually preferred (Turian et al., 2010).

Distributed representations provide a-priori knowledge to the input representation. The embeddings from such representations are dense and generalise easily. They serve as inputs for downstream NLP tasks. From a mathematical perspective, they serve as a dimensionality-reduction technique, where each dimension is a latent factor that encodes some information about the word (Mikolov et al., 2013b,a). They provide the advantage of some mitigation to the challenge of the *curse of dimensionality* (Bengio et al., 2003). Word2Vec is a shallow linear example of distributed representation. It trains quickly and has two arhcitectures for training: continuous Bag-of-Words (CBoW) and continuous Skip-gram, as depicted in Figure 3.1 (Mikolov et al., 2013b,a). Joulin et al. (2016) introduced fastText, which is an extension of word2vec. Subword vectors in fastText addressed morphology (the structure of words) by treating each word as the sum of a bag of character n-grams (Bojanowski et al., 2017), thereby addressing out-of-vocabulary (OOV) words by building vectors for words that are not in the training data (Bojanowski et al., 2017). The n-gram method differs and achieves less significant results when compared with the NN method (Bengio et al., 2003). Improving the results of NLP tasks using NN can involve the introduction of a-priori knowledge (Bengio et al., 2003). Such knowledge may include semantic information from WordNet and grammatical information from PoS. Indeed, the distributional context does not need to be textual alone. Texts are often illustrated with images and some approaches combine these, representing an image as a bag of keypoint features, giving rise to mixed visual and textual dimensions (Erk, 2012).

Table 3.2: Example of one-hot encoding

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
pat	1	0	0	0	0	0	0	0
let	0	1	0	0	0	0	0	0
the	0	0	1	0	0	0	0	0
cat	0	0	0	1	0	0	0	0
out	0	0	0	0	1	0	0	0
of	0	0	0	0	0	1	0	0
the	0	0	0	0	0	0	1	0
bag	0	0	0	0	0	0	0	1

Table 3.3: Example of low-dimensional, distributed representation

	1	**2**	**3**	**4**
pat	0.023	0.011	-0.013	0.201
let	0.11	-0.23	0.132	-0.221
the	0.312	0.033	0.078	0.091
cat	-0.165	0.099	0.076	0.045
out	0.088	0.109	0.076	0.023
of	0.156	-0.066	0.231	0.002
bag	0.002	0.014	-0.055	0.311

The continuous Skip-gram architecture selects pairs of target (or center) and context words and trains to predict whether the context word appears in the context window of the center word through an unsupervised process. An embedding layer is then added to serve as a lookup table. A similarity score that uses the dot product operator is calculated between the one-hot encoded context and center words. Negative sampling (Gutmann and Hyvärinen, 2012) is then applied, such that (*center, context*) pairs that do not occur in the sentences are assigned low similarity scores. The continuous Skip-gram is expressed in Equation 3.4 formally (Mikolov et al., 2013b), where the aim is to maximise the average log probability; the context size and center word are given by c and w_t, respectively. The other architecture, CBoW, considers simultaneously all words (or subwords) on both sides of the center word and trains to predict the center word (Mikolov et al., 2013a). The mean (or sum or any form of merger) of the context embedding is calculated and a softmax activation is attached for selecting the one-hot encoded context word (Mikolov et al., 2013b). The hierarchical softmax (Morin and Bengio, 2005) is an alternative function that may be applied, instead of negative sampling, to either of the architectures in word2vec. Additionally, subsampling of frequent words may be used to counter imbalance of rare and frequent words (Mikolov et al., 2013a). Another distributed representation: Glove, introduced by Pennington et al. (2014a), combines global matrix factorisation and local context window by training on non-zero elements of the co-occurrence matrix instead of the entire document.

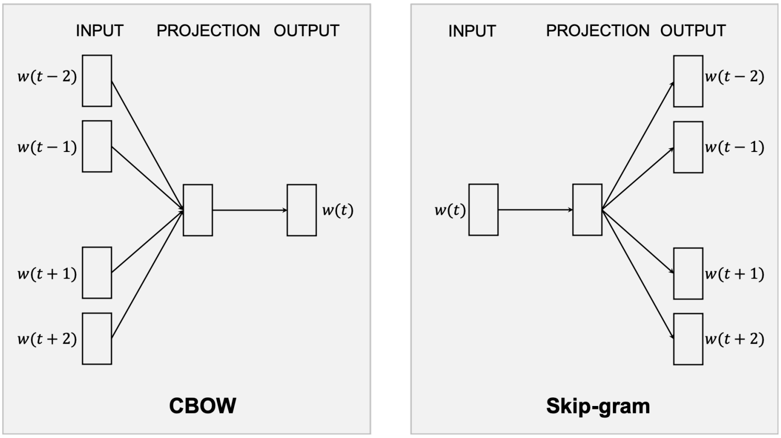

Figure 3.1: The CBoW and continuous Skip-gram model architectures (Mikolov et al., 2013a)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} log\ p(w_{t+j}|w_t) \tag{3.4}$$

3.3 Experiments & Evaluation: Shallow neural network (NN)

Levy et al. (2015) argued that choices about certain system design and hyperparameter optimisations are responsible for the differences that have been observed in the performance of word embeddings between NN-based and count-based (BoW) methods. This suggests that the choice of the combination of hyperparameters has significant impact on the performance of a given model. Also, Mikolov et al. (2013b) explained that the choice of hyperparameters is task-specific, as different tasks perform well under different combination of hyperparameters (Zhuang et al., 2021). The model architecture, the training window, subsampling rate and the dimension size of the vector were considered as the most important in their work. In order to explore the role of hyperparameters for word2vec embeddings, we conducted different sets of experiments with the following setup (Adewumi et al., 2022d). The Gensim (3.8.1) (Řehůřek and Sojka, 2010) Python (3.6.9) library implementation of word2vec was utilised to create word embeddings and to evaluate them on the analogy test sets. It should be noted that Faruqui et al. (2016) explains that there are problems with evaluation of embeddings by using word similarity tasks, which are part of analogy test. One of the problems is overfitting, which large datasets tend to alleviate (Stevens et al., 2020).

Multiple runs were conducted for some of the embeddings to validate if there is any significant difference in the evaluations between the runs, as it was prohibitively time-consuming to run every model multiple times. This is because the Python library takes several hours, on average, for most of the embeddings, given that it's an interpreted language (Adewumi, 2018). The Python implementation is slower than the original word2vec implementation. Raffel et al. (2020) made a similar assumption in their experiments because of the prohibitive cost of running experiments for each of their variant models multiple times. We extended work on embedding size to 3,000 dimensions and epochs of 5 and 10 (Adewumi et al., 2022d). Words with frequency less than 5 times in the datasets were dropped to form the vocabulary for the embeddings and stop words were also removed using the natural language toolkit (NLTK) (Bird et al., 2009).

In a second setup, the fastText original implementation in C++ was utilised (Grave et al., 2018). Although the programming language of this second setup was faster, the size of the datasets in this setup are still large, so a few hours were also needed to train each embedding. Hence, a similar approach in the first setup was adopted. The analogy test set by Mikolov et al. (2013b) is used to evaluate the embeddings, in a reasoning task, by running the evaluations in Gensim (3.8.1). It contains semantic and syntactic similarity tasks (Mikolov et al., 2013a). This is in addition to the WordSimilarity-353 (with Spearman correlation) by Finkelstein et al. (2002). The Swedish embeddings were evaluated using the same programs and the Swedish analogy test set (Adewumi et al., 2020c,b). Certain default hyperparameter settings were retained, as described by Bojanowski et al. (2017). In a third experimental setup (Adewumi et al., 2020b), involving the comparison of Swedish embeddings from two different corpora: the Swedish Wikipedia and the Gigaword corpora, the embeddings have 300 dimensions and are trained for 10 epochs.

Pytorch framework was used for the downstream tasks. As discussed in the previous chapter, data shuffling is carried out for the downstream tasks and the split ratio is 70:15:15 for the training, dev, and test sets, respectively. Multiple runs (four) per experiment are conducted and the averages taken. Given that a definite, useful evaluation of embeddings is best done when used for relevant downstream tasks (Chiu et al., 2016; Faruqui et al., 2016; Faruqui and Dyer, 2014; Lu et al., 2015; Gatt and Krahmer, 2018), two tasks are selected: NER and SA. The LSTM and the biLSTM are used for the tasks of NER and SA, respectively. These are depicted in Figures 3.2 and 3.3. The downstream experiments were run on a Tesla GPU on a shared DGX cluster running Ubuntu 18 while the embeddings are trained on a shared cluster running Ubuntu 16 with 32 CPU cores of 32x Intel Xeon 4110 at 2.1GHz. The biLSTM architecture includes an additional hidden linear layer before the output layer, when compared to the LSTM architecture that is used. Adam optimiser is utilised and a batch size of 64.

Datasets

The 2019 English Wiki news abstract of about 15M by Wikipedia (2019c), the 2019 English Simple Wiki (SW) articles of about 711M by Wikipedia (2019d) and the Billion Word (BW) corpus of 3.9G by Chelba et al. (2013) are used to train the models to create the embeddings in the first experimental setup. In other work, examples of training data that have been used in generating word embeddings include Google News (Mikolov et al., 2013a), Common Crawl, Gigaword (Mikolov et al., 2018; Pennington et al., 2014b) and Wikipedia (Bojanowski et al., 2017). The English Wikipedia in the second experimental setup is the 2019 Wikipedia dump of 27G (4.86B tokens) after preprocessing (Wikipedia, 2019a). The benchmark corpus, IMDB, by Maas et al. (2011) is used for SA. The original training set is what was available with the ground truth. from the data source. The set has 25,000 sentences with half having positive sentiments and the other half having negative sentiments. The Groningen Meaning Bank (GMB) by Bos et al. (2017) is used for NER. It contains 47,959 samples and 17 unique labels.

The Swedish Gigaword corpus that is used in the third experimental setup was generated as described by Rødven Eide et al. (2016) and the Wikipedia corpus was preprocessed using the script by Grave et al. (2018). The Gigaword corpus contains Wikipedia, among other sources, but appears to be limited to the science genre and year 2015 (Rødven Eide et al., 2016). The Wikipedia corpus that is compared in this experimental setup is the full version (containing all genres), serving as a kind of ablation study. It covers topics, including those of the Swedish Gigaword corpus, and in addition, entertainment, art, politics, and more, and spans several years. The recommended script that is used to preprocess the Wikipedia corpus returned all text as lowercase and did not retain non-ascii characters, which distorted some of the Swedish words. Apparently, the script is only best for English data. Despite this noise in the preprocessed data, a portion of it was tested for coherence on Google Translate and the English translation returned was largely meaningful. It appears the noise issue was not serious enough to adversely affect the models created. A better alternative, however, would have been to test the Swedish corpus as is (despite portions of English content) or use another Swedish Wikipedia cor-

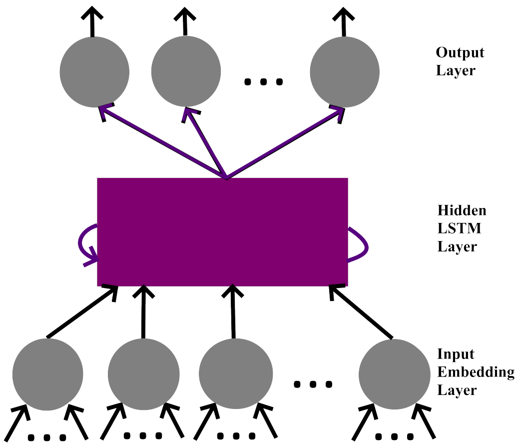

Figure 3.2: Network architecture for NER

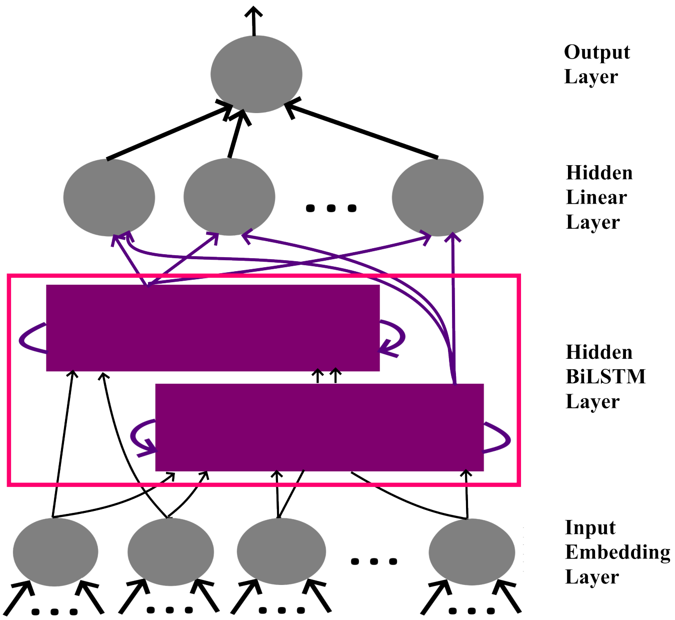

Figure 3.3: Network architecture for SA (Adewumi et al., 2022d).

pus that retained the peculiarities of the language, even after preprocessing. Hengchen and Tahmasebi (2021) produced such data at a later point when they introduced the

Supersim evaluation dataset for Swedish. The Gigaword corpus has a file size of 5.9G and contains 1.08B tokens while the Swedish Wikipedia has a file size of 4.2G and contains 767M tokens (Wikipedia, 2019b). They were pre-processed using the recommended script by Grave et al. (2018).

The cleaned 2020 Yorùbá Wikipedia dump (182M) (Wikipedia, 2020) containing diacritics (tonal marks) to different degrees across the articles and a normalised (undiacritised) version is used in the fourth experimental setup (Adewumi et al., 2020a). In addtion, the largest, diacritised data used by Alabi et al. (2020) is used to compare the performance of embeddings in this work. The original Yorùbá Wikipedia dump was very unsuitable for training and required large manual cleanup. We also created two Yorùbá analogy test sets: one with diacritics and an exact copy without diacritics Adewumi et al. (2020a). Evaluation is done with only the diacritised version of the analogy set and the Yorùbá WordSim by Alabi et al. (2020). Performance on the Yorùbá analogy test sets were very poor and may not be very important.

3.3.1 Hyperparameter exploration for word2vec

We chose grid search to explore the hyperparameters, based on the literature (Mikolov et al., 2013b). The hyperparameters are given in Table 3.4. Eighty runs per dataset were conducted for the Wiki news abstract and the Simple Wiki. Experiments for all combinations for 300 dimensions were conducted on the Billion Word corpus, plus additional runs for the window size 8 + Skip-gram (s1) + hierarchical softmax (h1) combination. This is to establish the behaviour of quality of word vectors as dimensions are increased without increasing the data size. Table 3.5 shows the hyperparameter choices for the two networks for the downstream tasks. The metrics for extrinsic evaluation include F1, precision, recall and accuracy (for SA).

Table 3.4: Embeddings hyperparameter choices (Adewumi et al., 2022d). (notations based on Gensim library convention)

Hyper-parameter	Values
Dimension size	300, 1200, 1800, 2400, 3000
Window size (w)	4, 8
Architecture	Skipgram (s1), CBoW (s0)
Algorithm	H. Softmax (h1), N. Sampling (h0)
Epochs	5, 10

Results show a major advantage of training with relatively smaller corpora, as depicted in Table 3.6. The training time and average loading time for our embeddings into the downstream model are considerably shorter. This is representative of similar embeddings. The Gensim WordSim output file always has more than one evaluation score reported, including the Spearman correlation, as given in Table 3.7. The first value from the program is a cosine similarity variant and is reported as WordSim score1 in

Table 3.5: Downstream network hyperparameters (Adewumi et al., 2022d).

Archi	Epochs	Hidden Dim	LR	Loss
LSTM	40	128	0.01	Cross Entropy
BiLSTM	20	128 * 2	0.0001	BCELoss

the above-mentioned table. It summarises results from the intrinsic evaluations for 300 dimensions. The smallest dataset (Wiki news abstract) results are so poor that they are not required. This outcome should be because of the tiny file size (15M).

Table 3.6: Embedding training & loading time (Adewumi et al., 2022d). (w: window size, s1: skipgram, h1: hierarchical softmax, h0: negative sampling)

Model	Training (hours)	Loading Time (s)
SW w8s1h0	5.44	1.93
BW w8s1h1	27.22	4.89
GoogleNews (Mikolov et al., 2013a)	NA	97.73

As can be observed from Table 3.7, the combination of skipgram-negative sampling (s1h0) generally performs better. The embedding by Mikolov et al. (2013a) achieves the highest analogy score, however, the skipgram-negative sampling embedding of window size 8 of the SW achieves the highest WordSim score1 and Spearman correlation. It is noteworthy that the GoogleNews embedding is based on a vocabulary size of 3M, a large figure when compared to recent SoTA embeddings (Devlin et al., 2018a). The SW has a vocabulary size of 368K while the BW has 469K. Figure 3.4 gives similar trend for the two datasets depicted, SW and BW, where scores improve but start to drop after over 300 dimensions. This observation is true for all the combinations and is also confirmed by Mikolov et al. (2013a).

For the downstream tasks, comparable performance in accuracy is achieved in SA to that by Maas et al. (2011), though less than half of the dataset for training is used. Notably also, evaluation is on a smaller different size. Tables 3.8 and 3.9 summarise key results for the NER and SA tasks, respectively. The BW Skip-gram-negative sampling (w4s1h0) embedding performs best in F1 score for the NER task. Interestingly, the same embedding has the best analogy score among the models generated. The default Pytorch embedding trails behind most of the pretrained embeddings by a small amount. However, it outperforms the pretrained embeddings in accuracy and F1 scores in the SA task. The CBoW-negative sampling of the SW performs relatively well in both the downstream tasks. For power of 1 and alpha of 0.05, significance tests of the difference of two means of the two-sample t-test for the F1 scores give p-values < 0.0001 in the two cases, i.e., the 100B and the skipgram-negative sampling (w4s1h0) of the BW embedding for NER, and the CBoW-negative sampling (w8s0h0) for the SW for SA.

Table 3.7: Scores for 300 dimensions for 10 epochs for SW, BW & GoogleNews corpora (Adewumi et al., 2022d). (w: window size, s1: skipgram, s0: CBoW, h1: hierarchical softmax, h0: negative sampling / notations are based on the Gensim convention)

	w8s1h1	w8s0h1	w8s0h0	w8s1h0	w4s1h1	w4s0h1	w4s0h0	w4s1h0
Simple Wiki (SW)								
Analogy	0.461	0.269	**0.502**	0.439	0.446	0.243	0.478	0.407
WordSim score1	0.636	0.611	0.654	**0.655**	0.635	0.608	0.620	0.635
Spearman	0.670	0.648	0.667	**0.695**	0.668	0.648	0.629	0.682
Billion Word (BW)								
Analogy	0.587	0.376	0.638	0.681	0.556	0.363	0.629	**0.684**
WordSim score1	0.614	0.511	0.599	**0.644**	0.593	0.508	0.597	0.635
Spearman	0.653	0.535	0.618	**0.681**	0.629	0.527	0.615	0.677
GoogleNews - 100B (s1h0)								
Analogy: 0.740			WordSim score1: 0.624			Spearman: 0.659		

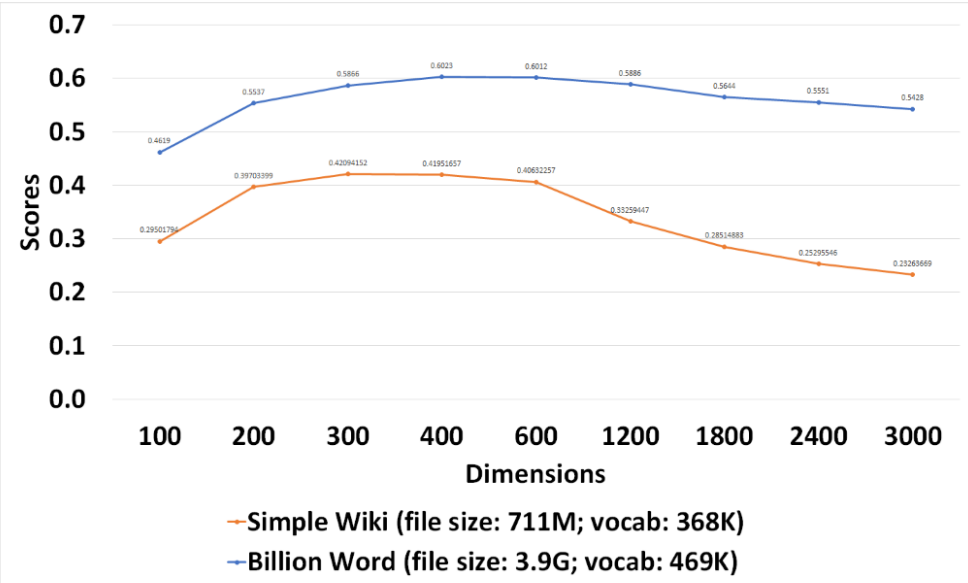

Figure 3.4: Analogy Scores for Skip-gram-hierarchical softmax (w4s1h1) of SW for 5 Epochs & Skip-gram-hierarchical softmax (w8s1h1) of BW for 10 epochs (Adewumi et al., 2022d). (not drawn to scale from 400)

3.3.2 Swedish embeddings and the analogy set

Section 2.2 discusses the Swedish analogy test set in detail. The unavailability of such a set (Fallgren et al., 2016; Précenth, 2019), which is similar to the English set by Mikolov et al. (2013b), motivated the creation of one (Adewumi et al., 2020c). From

Table 3.8: NER Dev & Test sets Mean Results (Adewumi et al., 2022d). (w: window size, s1: skipgram, s0: CBoW, h0: negative sampling)

Metric	**Default**	**100B**	**w8 s0 h0**	**w8 s1 h0**	**BW w4 s1 h0**
	Dev, Test	Dev, Test	Dev, Test	Dev, Test	Dev, Test
F1	0.661, 0.661	**0.679**, 0.676	0.668, 0.669	0.583, 0.676	**0.679, 0.677**
Precision	0.609, 0.608	**0.646, 0.642**	0.636, 0.637	0.553, 0.642	0.644, **0.642**
Recall	0.723, **0.724**	0.716, 0.714	0.704, 0.706	0.618, 0.715	0.717, 0.717

Table 3.9: SA Dev & Test sets Mean Results (Adewumi et al., 2022d). (w: window size, s1: skipgram, s0: CBoW, h0: negative sampling)

Metric	**Default**	**100B**	**w8 s0 h0**	**w8 s1 h0**	**BW w4 s1 h0**
	Dev, Test	Dev, Test	Dev, Test	Dev, Test	Dev, Test
F1	**0.810, 0.805**	0.384, 0.386	0.798, 0.799	0.548, 0.553	0.498, 0.390
Precision	0.805, 0.795	0.6, 0.603	**0.814, 0.811**	0.510, 0.524	0.535, 0.533
Recall	**0.818, 0.816**	0.303, 0.303	0.788, 0.792	0.717, 0.723	0.592, 0.386
Accuracy	**0.807, 0.804**	0.549, 0.55	0.801, 0.802	0.519, 0.522	0.519, 0.517

Tables 3.10 and 3.11, we notice the good performance of the Skip-gram-negative sampling embeddings in all (English & Swedish) but one case. Again, this confirms previous work (Mikolov et al., 2013a). Notable is the higher performance of the CBoW-negative sampling embedding compared to the GoogleNews-based embedding by Mikolov et al. (2013a), though the earlier is from a smaller corpus. The subword embedding by Grave et al. (2018) has the highest performance overall.

Table 3.10: Skip-gram English & Swedish intrinsic scores (Adewumi et al., 2020c). (highest score in bold). H.S.: hierarchical softmax; N. S.: negative sampling

	Skip-gram (s1)			
	H. S. (h1)		**N. S. (h0)**	
window (w)	**4**	**8**	**4**	**8**
Subword %				
Analogy	62.6	58.8	74.4	69.8
WordSim score1	64.8	66.3	69.9	**70**
Spearman	67.6	69.4	**74.3**	73.6
Word2Vec %				
Analogy	61.3	58.3	73.5	70.4
WordSim score1	66.3	67.3	69.6	**70.1**
Spearman	70	70.9	74.5	**74.7**
Swedish				
Subword %	45.05	39.99	53.53	53.36
Word2Vec %	45.53	41.21	**58.25**	57.30

Table 3.11: CBoW English & Swedish intrinsic scores (Adewumi et al., 2020c). (highest score in bold). H.S.: hierarchical softmax; N. S.: negative sampling; Gr: (Grave et al., 2018), GN: Google News (Mikolov et al., 2013a)

	CBoW (s0)					
	H. S. (h1)		**N. S. (h0)**		**Gr**	**GN**
window (w)	**4**	**8**	**4**	**8**		
Subword %						
Analogy	67.2	68.7	71.6	71	**82.6**	
WordSim score1	62.6	66.2	47.3	51.1	68.5	
Spearman	65.3	70.3	45.3	49.5	70.2	
Word2Vec %						
Analogy	59.7	61.9	**76.2**	75.4		74
WordSim score1	64.1	66.7	65.4	67.5		62.4
Spearman	68.2	71.2	66.9	69.4		65.9
Swedish						
Subword %	26.5	23.93	36.79	35.89	**60.9**	
Word2Vec %	28.02	28.04	52.81	55.64		

The results from comparing the Swedish Gigaword and Wikipedia corpora are presented in Tables 3.12 and 3.13 for the initial learning rates of 0.05 and 0.01, respectively. The Skip-gram-negative sampling combination for both corpora for subword and word2vec models have the best scores in most cases. No value is recorded for the Gigaword CBoW-hierarchical softmax with the initial LR of 0.05 because the program fails several times, as it reports *Encountered NaN* error. The highest score (of 60.38%) from both tables belongs to the word2vec embedding of the Wikipedia corpus while the lowest (of 2.59%), belongs to the CBoW-hierarchical softmax, subword embedding of the Gigaword corpus. The better performance of the Wikipedia corpus, despite its noise, we conjecture may be due to the wider genre/topic coverage (or balance in domains), the relatively small noise in the corpus or the combination of both. Nearest neighbour qualitative assessment of the Skip-gram-negative sampling subword embedding is presented in Section 3.14.

We further show in experiments with the Yorùbá language that embeddings from the undiacritised Wikipedia (U_Wiki) outperforms C3 by Alabi et al. (2020), giving the highest WordSim score & corresponding Spearman correlation, as indicated in Tables 3.15 and 3.16 (Adewumi et al., 2020a). Wiki, U_Wiki, C3 & CC, represent embeddings from the cleaned Wikipedia dump, the undiacritised version, the diacritised data from Alabi et al. (2020), and the Common Crawl embedding by Grave et al. (2018), respectively. The negative effect of noise, from the original data, in the Wiki word2vec embedding appears to reduce in the subword version in Table 3.16.

Table 3.12: Mean Analogy Scores for Swedish Gigaword & Wikipedia Corpora with LR=0.05 (Adewumi et al., 2020b).

	Skipgram (s1)				**CBoW (s0)**			
	H. S. (h1)		**N. S. (h0)**		**H. S. (h1)**		**N. S. (h0)**	
window (w)	**4**	**8**	**4**	**8**	**4**	**8**	**4**	**8**
Word2Vec %								
Wikipedia	47.02	44.09	**60.38**	60.38	29.09	30.09	54.39	56.81
Gigaword	40.26	44.23	**55.79**	55.21	26.23	27.82	55.2	55.81
Subword %								
Wikipedia	46.65	45.8	**56.51**	56.36	28.07	24.95	38.26	35.92
Gigaword	41.37	44.7	**58.31**	56.28	2.59	-	46.81	46.39

Table 3.13: Analogy Scores for Swedish Gigaword & Wikipedia Corpora with LR=0.01 (Adewumi et al., 2020b).

	Skipgram (s1)				**CBoW (s0)**			
	H. S. (h1)		**N. S. (h0)**		**H. S. (h1)**		**N. S. (h0)**	
window (w)	**4**	**8**	**4**	**8**	**4**	**8**	**4**	**8**
Word2Vec %								
Wikipedia	48.92	49.01	51.71	**53.48**	32.36	33.92	47.05	49.76
Gigaword	39.12	43.06	48.32	**49.96**	28.89	31.19	44.91	48.02
Subword %								
Wikipedia	45.16	**46.82**	35.91	43.26	22.36	21.1	14.31	14.45
Gigaword	39.13	43.65	45.51	**49.1**	31.67	35.07	28.34	28.38

3.4 Contextual vs non-contextual representation

Researchers have criticised the VSM (Budanitsky and Hirst, 2001; French and Labiouse, 2002; Turney and Pantel, 2010). The main criticism is that they largely ignore word order; for example, the words *rock* and *solid* in "rock solid" and "solid rock" will be individually represented with the same vectors, even though the phrases are different in meaning (Turney and Pantel, 2010). Another problem, according to French and Labiouse (2002), is the absence of essential world knowledge. Contextual word representations, which are derived from deep bidirectional language model (LM) have demonstrated significant improvement by capturing contextual semantic structures that outperform word embeddings, thereby improving the SoTA (Peters et al., 2018b). They differ from non-contextual embeddings because each token is assigned a representation that is a function of the entire input sentence, instead of a context window (Peters et al., 2018a). An LM is a probability distribution over a sequence of tokens (Liu et al., 2020a). Language models that achieve SoTA first produce context-insensitive token representation, producing an embedding lookup. Context-dependent representations are then computed afterwards (Peters et al., 2018b).

There are many models that use contextual representation. Embeddings from Lan-

Table 3.14: Example qualitative assessment of Swedish Skip-gram-negative sampling (w4s1h0) subword embedding (Adewumi et al., 2020b).

Nearest Neighbor	**Result**
Wiki: syster	systerdotter (0.8521), systern (0.8359), ..
Gigaword: syster	systerdotter (0.8321), systerdottern (0.8021), ..

Table 3.15: Yorùbá word2vec embeddings intrinsic scores (%) (Adewumi et al., 2020a).

Data	Vocab	Analogy	WordSim	Spearman
Wiki	275,356	0.65	26.0	24.36
U_Wiki	269,915	**0.8**	**86.79**	**90**
C3	31,412	0.73	37.77	37.83

Table 3.16: Yorùbá subword embeddings intrinsic scores (%) (Adewumi et al., 2020b).

Data	Vocab	Analogy	WordSim	Spearman
Wiki	275,356	0	45.95	44.79
U_Wiki	269,915	0	**72.65**	**60**
C3	31,412	0.18	39.26	38.69
CC	151,125	4.87	16.02	9.66

guage Models (ELMo) (Peters et al., 2018a), Generative Pre-trained Transformer (GPT)-2, and Text-to-Text Transfer Transformer (T5) (Raffel et al., 2019), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018a), and its many successors like Robustly optimized BERT pretraining Approach (RoBERTa) (Zhuang et al., 2021) are just some of them. Specifically, ELMo is a deep contextualised representation that models complex (syntactic and semantic) characteristics of word use, and how they vary in different contexts. ELMo representations are a function of all of the internal layers of the biLM, making them deep, just as it is with BERT and many recent SoTA models. Usually, in these pretrained contextualised models, the higher-level states (or upper layers) of the model capture context-dependent aspects of word meaning while lower-level states model aspects of syntax. Simultaneously exposing all of these signals is highly beneficial (Peters et al., 2018a; Devlin et al., 2018a). Compared to BERT, ELMo might be considered shallow. The BERT model, which is based on the encoder stack of the Transformer architecture, is a bidirectional pretrained model from unlabeled text. The Transformer is an encoder-decoder architecture based solely on the attention mechanism (Vaswani et al., 2017). Its architecture is depicted in Figure 3.5. BERT was pretrained by jointly conditioning on the left and right context in all the layers of the model (Devlin et al., 2018a). It is based on the WordPiece embedding. The input representation for a token is constructed by summing the corresponding token, segment, and position embeddings. The depiction is given in Figure 3.6. (Devlin et al., 2018a)

The encoder of the original Transformer has a stack of 6 identical layers, with 2 sub-layers in each. A mulit-head self-attention and a fully connected feed-forward network occupy the first and second sub-layers, respectively. Additional structures complete the

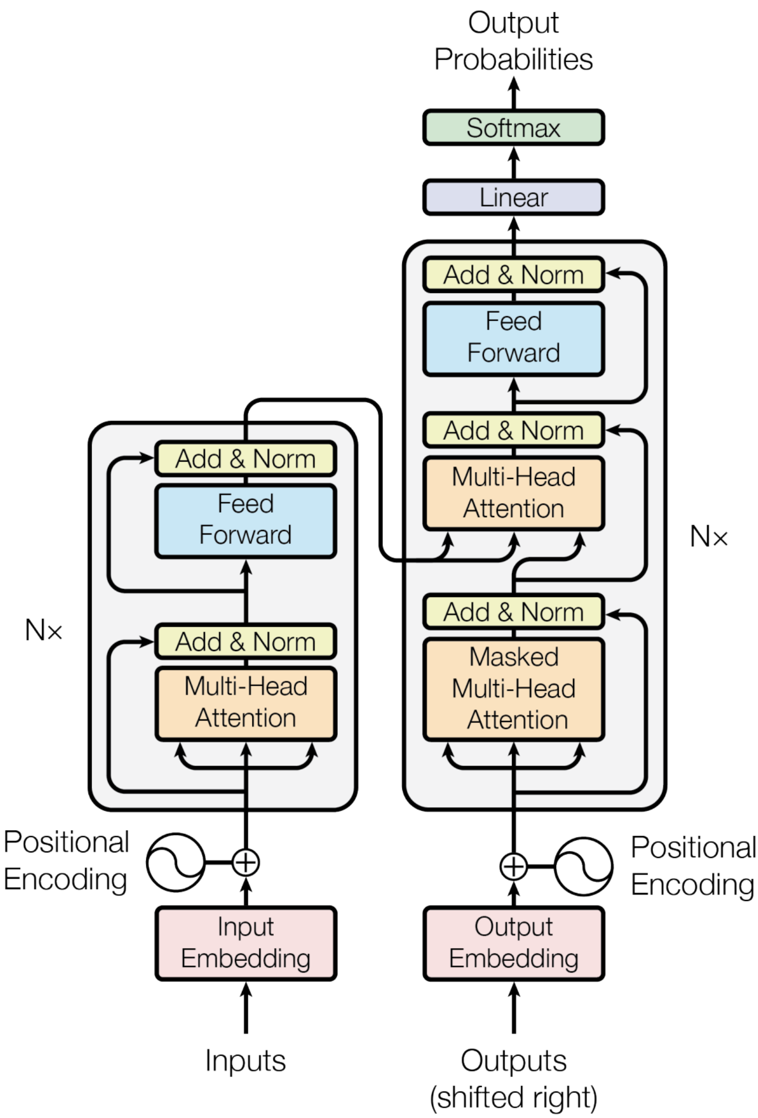

Figure 3.5: The Transformer architecture by Vaswani et al. (2017)

encoder. The decoder is very similar to the encoder but it has a third sub-layer that performs multi-head attention over the output from the encoder. Masking is added to the first sub-layer's attention to prevent positions from attending to subsequent position (Vaswani et al., 2017). Positional encoding to the input is needed by the Transformer

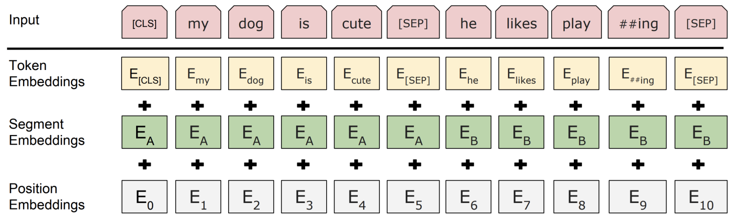

Figure 3.6: BERT input representations, which are the sum of the token embeddings, the segmentation embeddings, and the position embeddings (Devlin et al., 2018a).

at the initial points of both the encoder and decoder stacks because the model has no recurrence or convolution, which are useful for the order of sequence of input (Vaswani et al., 2017).

Three common encoding algorithms in recent SoTA LM are WordPiece (Schuster and Nakajima, 2012), BPE (Gage, 1994; Sennrich et al., 2016), and sentencepiece (Kudo and Richardson, 2018). WordPiece is similar to BPE and sentencepiece incorporates BPE. WordPiece initialises the vocabulary so that it includes all the characters present in the training data and learns a number of merge rules progressively (Schuster and Nakajima, 2012). It runs a greedy algorithm and chooses the symbol pair that maximises the likelihood of the training data in the vocabulary. BPE compresses by segmenting rare words into more commonly appearing subwords. Common pairs of adjacent bytes are replaced by single bytes that is not in the original data. The process is repeated until there is no further compression possible. Its expansion routine is fast and it's not memory intensive, usually. The original algorithm was unable to handle large files that are too big to fit into memory (Gage, 1994; Radford et al., 2019). Sennrich et al. (2016) introduced an improvement to the original BPE algorithm. Instead of merging pairs of bytes, they merge characters, thereby encoding rare or unknown words as sequences of subword units. Each word is represented as a sequence of characters. This version of BPE is used in sentencepiece. Sentencepiece is an unsupervised tokeniser/detokeniser for text-generation NN systems, such as DialoGPT, where the size of the vocabulary is determined before training (Kudo and Richardson, 2018).

Using contextual embeddings

Contextual embeddings are useful for downstream tasks in NLP. The three main ways they may be used are feature-extraction methods, finetuning, and adapter methods (Liu et al., 2020a). ELMo is based on feature-extraction. It freezes the weights and forms a linear combination of the representations, which is then used as features for task-specific architectures (Liu et al., 2020a). Peters et al. (2018a) found that using ELMo at the output of the model, besides being input at the initial layer, in task-specific architectures improves results in some tasks. Devlin et al. (2018a) also compared this approach to finetuning by supplying the contextual embeddings to a randomly initialised two-layer biLSTM bfore the classification layer. Finetuning starts with the pretrained contextual weights of the model and makes small adjustments to them to specialise them to specific tasks. Usually, a linear layer is added on top of the pretrained model in the finetuning process (Devlin et al., 2018a; Liu et al., 2020a). Adapters are modules added between layers of a pretrained model, whose weights are fixed, with a multi-task learning objective (Houlsby et al., 2019; Liu et al., 2020a; Raffel et al., 2020). The adapter modules are tuned, adding only few parameters per task, unlike the usual 100% weight adjustment for finetuning.

3.5 Experiments & Evaluation: Named Entity Recognition (NER) for African languages

We investigated the performance of deep NNs for NER on various low-resource African languages (Adelani et al., 2021). The languages are Ahmaric, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian-Pidgin English, Swahili, Wolof, and Yorùbá. Some of the languages are further discussed in Section 5.1. The languages were selected primarily because of the availability and willingness of collaborators who annotated data. Characteristics of some of the languages that could pose challenges for systems developed for English include diacritics (or tonal marks) and the use of non-latin characters. The experimental setup for NER for the ten languages involved sourcing data from online news websites and recruiting collaborators to annotate the data. There were 2 to 6 annotators/language, who are native/L1 speakers, and IAA is calculated per language. Each language has about 2,500 labelled sentences on average. The Hausa language, from Table 3.17, had the best F1 score of 91.64 and the XLM-R has the best overall performance as a model.

*Table 3.17: Transfer Learning average F1 Results over 5 runs. 3 Tags: `PER`, `ORG` & `LOC`. WikiAnn, `eng`-CoNLL, and the annotated datasets are trained for 50 epochs while fine-tuning is for 10 epochs. Highest score/language is in **bold**, and the best score in zero-shot setting is indicated with an asterisk (*) (Adelani et al., 2021).*

Method	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor
XLM-R-base	69.71	91.03	86.16	73.76	80.51	75.81	86.87	**88.65**	69.56	78.05
WikiAnn zero-shot	27.68	–	21.90	9.56	–	–	–	36.91	–	10.42
eng-CoNLL zero-shot	–	67.52	47.71	38.17	39.45	34.19	67.27	76.40	24.33	39.04
pcm zero-shot	–	63.71	42.69	40.99	43.50	33.12	–	72.84	25.37	35.16
swa zero-shot	–	85.35*	55.37	58.44	57.65*	42.88*	72.87*	–	41.70	57.87*
hau zero-shot	–	–	58.41*	59.10*	59.78	42.81	70.74	83.19*	42.81*	55.97
WikiAnn + finetune	**70.92**	–	85.24	72.84	–	–	–	87.90	–	76.78
eng-CoNLL + finetune	–	89.73	85.10	71.55	77.34	73.92	84.05	87.59	68.11	75.77
pcm + finetune	–	90.78	86.42	71.69	79.72	75.56	–	87.62	67.21	78.29
swa + finetune	–	91.50	87.11	74.84	80.21	74.49	86.74	–	68.47	**80.68**
hau + finetune	–	–	86.84	74.22	80.56	75.55	88.03	87.92	**70.20**	79.44
combined East Langs.	–	–	–	**75.65**	81.10	77.56	–	88.15	–	–
combined West Langs.	–	90.88	87.06	–	–	–	87.21	–	69.70	**80.68**
combined 9 Langs.	–	**91.64**	**87.94**	75.46	**81.29**	**78.12**	**88.12**	88.10	69.84	80.59

The models trained are CNN-biLSTM-CRF, mBERT, and XLM-R. The latter two models are based on pretrained models from the HuggingFace hub (Wolf et al., 2020). Additional techniques employed in the study involves combining XLM-R and gazetteers, cross-lingual transfer learning (from English, using the CoNLL-2003 dataset by Aggarwal and Zhai (2012), and Swahili), and the use of the cross-lingual WikiAnn dataset (Pan et al., 2017). A gazetteer is an index that typically contains geographical information (or place-names) and social statistics and is used in conjunction with a map (Grover and Tobin, 2014). Language-specific finetuning of BERT and XLM-R on unlabelled data is also done for each of the languages, thereby providing additional performance improvements when compared with mainly finetuning mBERT and XLM-R, respectively. It was observed from the study that the pretrained models have reasonable performance on languages that they were not pretrained on but showed better performance if the language was part of the pretrained languages. Also, across all the languages, it is observed that entities that were not in the training data and those which are three-word entities or more were challenging for the models.

Chapter 4

Open-Domain Conversational Systems

*"**Garbage in, garbage out**, that's the way codes go."*

(Parallelism)

In the West African folktale by Medearis (1995), objects like yam, cloth, water, and a royal throne spoke to humans. The people to whom the objects spoke were so shocked that they nearly jumped out of their skin. This story might be unrealistic. However, metal boxes or handheld devices having conversations with humans is becoming more ubiquitous. Conversational systems may be classified on the basis of architecture into frame-based, rule-based and data-driven approaches (Jurafsky and Martin, 2020). They may also be classified on the basis of their goal into task-based and open-domain approaches (Hosseini-Asl et al., 2020).

This chapter is organised as follows. First, Section 4.1 discusses the characteristics of human dialogues before discussing open-domain versus task-based conversational systems in Section 4.2. Deep models for open-domain conversational systems are introduced in Section 4.3 before looking at measuring progress of conversational systems in Section 4.4. The following Sections 4.5 and 4.6 then take a look at metaphors in chatbots and experiments & evaluation, respectively, before closing with ethics of developing conversational systems.

4.1 Characteristics of human dialogues

Human dialogues can be complex (Jurafsky and Martin, 2020). We do not only converse using speech but we use gestures and facial expressions, usually called body language. Even when we write during conversations, we may employ cues such as confirmatory/clarification questions or mimick sound in what is called *onomatopoeia*. Clarification questions for confirmation are particularly useful in task-based systems before filling slots or deciding intents (Jurafsky and Martin, 2020). An example of human-human conversation from the training set of the MultiWOZ dataset is shown in Figure 4.1. The conversation covers the domains of booking a hotel. It shows turns of a conversation, where a turn is each single contribution to the conversation from a speaker (Schegloff, 1968; Jurafsky and Martin, 2020). There are a total of 10 turns in the figure. It will be observed that a turn can have more than one sentence. The turns may also be called utterances or dialogue acts (Jurafsky and Martin, 2020).

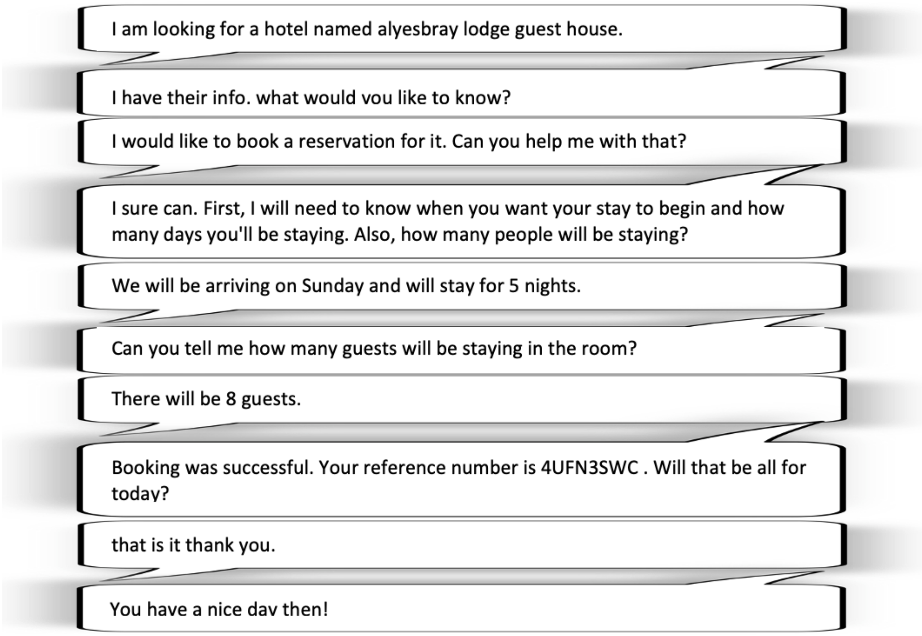

Figure 4.1: Conversation from the training set of the MultiWOZ dataset

Humans learn over time when the other converser (or speaker) in a dialogue may have paused, stopped (called endpointing) or might be making a correction (Jurafsky and Martin, 2020). Grounding is the useful feedback that one party in a conversation understood the other's utterance. It is how humans acknowledge the other party's utterance in a conversation. In human-human conversation, grounding may be indicated by "ok" or "I see" in responses by the hearer. Conversational systems need to understand these also. For example, in Figure 4.1, the 4th turn in the conversation responds to the first speaker with "I sure can. First, ...". The first sentence in the response is the grounding that indicates to the first speaker that the second speaker understood the request for reservation.

It is important to realise that a conversation is not a collection of independent turns but connected utterances. An exemption to this intuition was made when the assumption for training on the PIE-English idioms corpus was introduced and the corpus is used for training conversational models in Section 4.6. This assumption holds, in this case, because the sentences of the turns in the dataset discuss the same cases of idioms (or "domain"), even though the sentences are drawn from different examples from the base corpora: the BNC[1] and UKWaC (Ferraresi et al., 2008). This is further discussed in Section 4.5. Good examples of connected utterances are adjacency pairs, which are composed of first and second pair parts (Sacks et al., 1978). Examples of adjacency pairs are question-answer turns, compliment-appreciation turns, and proposal-acceptance turns. Furthermore, conversations do not always follow a predefined manner; side sequence or sub-dialogue within an ongoing dialogue may arise (Jurafsky and Martin, 2020). In addition, humans may also introduce new topics (or domain) in an ongoing conversation, which may change the direction of the conversation altogether. According to Sacks et al. (1978), the following are some of the observations in any human conversation.

- One party talks per time.
- Turn order varies.
- Turn size varies (Schegloff, 1968).
- Recurring change of speaker. This is when conversers alternate their roles between listening and speaking.
- Length of conversation is not known in advance.
- The number of participants can vary.
- Turn-allocation techniques may be used.
- Turn-taking errors may be fixed through helpful mechanisms, such as pausing for the next speaker.

One party may have the conversational initiative in a dialogue. This is the case when such a party controls the conversation. An example of this is an interview where the speaker asking the questions directs the conversation. This is the style for QA dialogue systems. However, in a typical human-human conversation, the initiative shifts back and forth between parties. Mixed (or rotating) initiative is harder to achieve than when one side controls the initiative in conversational systems. Designing them as passive responders is much easier (Jurafsky and Martin, 2020).

[1] english-corpora.org/bnc

4.2 Open-domain vs Task-based

A task is a specific piece of work to be accomplished[2]. Multi-task, therefore, implies multiple tasks are involved. Open-domain conversation refers to the unrestrained coverage of the topics of conversation (i.e. conversation around many domains or tasks) (Hosseini-Asl et al., 2020). The topics of conversation for humans can be many and varied at social events. Task-based (single-domain or closed-domain) systems tend to be rule-based (Jurafsky and Martin, 2020). Understanding input, deciding actions, and generating a response are usually the processes involved in task-based conversational systems (Hosseini-Asl et al., 2020). These processes are similar to what obtains with the NLU and NLG of open-domain conversational systems (Gehrmann et al., 2021). ELIZA by Weizenbaum (1969) is an example of a rule-based system. There are other examples of rule-based systems, such as PARRY (Colby et al., 1971). Such systems are designed with if-else conditions. Research systems (which are rule-based) consist of hand-crafted semantic grammars with thousands of rules (Jurafsky and Martin, 2020). The semantic grammar is a context-free grammar. The rule-based approach is popular in industry and has the advantage of high precision, however, the rules can be expensive, slow to create, and suffer from recall problems (Chowdhary, 2020; Jurafsky and Martin, 2020).

Since open-domain conversational systems are usually data-intensive, deep ANNs are more suitable than rule-based architectures, according to Jurafsky and Martin (2020). More is discussed about some of the architectures for open-domain conversational systems or NLG in Section 4.3. Data-driven systems learn inductively from large datasets of samples of conversations. The data available for such systems include transcripts of human-human spoken dialogues, such as the Gothenburg Dialogue Corpus (GDC) (Allwood et al., 2003), written dialogues, such as the MultiWOZ (Eric et al., 2020), crowdsourced conversations that are written, such as the EmpatheticDialogues (Rashkin et al., 2019), and social media conversations, such as Reddit[3]. Since the amount of data needed for training deep models is generally large, models are usually pretrained on large, unstructured text or conversations from social media before they are finetuned on specific conversational data, through transfer learning. The data-driven approach may be combined with the rule-based approach in a hybrid setting (Jurafsky and Martin, 2020).

4.2.1 Information Retrieval (IR)

One of the two common ways that data-driven conversational systems produce turns as response is through Information Retrieval (IR) (Jurafsky and Martin, 2020), where the system fetches information from some fitting corpus, given a dialogue context. Incorporating ranking and retrieval capabilities provides additional possibilities for chatbot response generation. If D is the training set of conversations, given a context (or query) q, the goal is to retrieve an appropriate turn r as the response. Similarity is used as the scoring metric and the highest scoring turn in D is selected from a potential set. This

[2]dictionary.cambridge.org
[3]reddit.com

may be achieved using different IR methods, including the classic tf-idf for D and q, and choosing the response with the highest cosine similarity with q (Jurafsky and Martin, 2020). This is expressed in Equation 4.1. A neural IR method is another approach one could use. For example, in an encoder-encoder architecture, one could train the first encoder to encode the query while the second encoder encodes the candidate response and the score is the dot product between the two vectors from both encoders.

NER facilitates IR, of which Information Extraction (IE) is a subtask (Aggarwal and Zhai, 2012). It is a main subtask of Information Extraction (IE) and uses tagging and partial parsing to identify (real-world) entities of interest (Aggarwal and Zhai, 2012; Indurkhya and Damerau, 2010). These entities are categories that include proper or special names, such as person, location, organization, date, time, money, percent, facility, and geo-political entities (Bird et al., 2009; Indurkhya and Damerau, 2010). The other main subtask of IE is relation extraction (Aggarwal and Zhai, 2012). IE derives meaning by building structured data from unstructured data. One method is to use triples to establish the meaningful relationships (Bird et al., 2009).

$$response(q,D) = \arg\max_{r \epsilon D} \frac{q.r}{|q||r|} \tag{4.1}$$

4.2.2 Natural Language Generation (NLG)

The other common method for turns as response for data-driven conversational systems is generation (Jurafsky and Martin, 2020). In this method, an encoder-decoder or language model is used for response generation, given a dialogue context. As shown in Equation 4.2, each token of the response (r_t) of the encoder-decoder model is generated by conditioning on the encoding of the query (q) and all the previous responses ($r_{t-1}...r_1$), where w is a word in the vocabulary V.

$$r_t = \arg\max_{w \epsilon V} P(w|q, r_{t-1}...r_1) \tag{4.2}$$

Decoding algorithms

The choice of the decoding algorithm in the encoder-decoder or decoder-only (autoregressive) models has a major impact on the performance of the model and the quality of responses that are generated (Holtzman et al., 2020). The random algorithm is a stochastic decoding method. The greedy algorithm has a tendency to produce repetitive and predictable tokens that lead to poor performance and beam search algorithm fairs better than it (Holtzman et al., 2020; Radford et al., 2019; Raffel et al., 2020). Beam search uses depth-first search and maintains the top k candidates on a priority queue for exploration. Both search algorithms are sometimes referred to as maximisation-based algorithms (Holtzman et al., 2020). Nucleus (or Top-p) sampling samples from the dynamic nucleus of tokens with the majority of the probability mass, cutting off the tail of the distribution that is deemed unreliable (Holtzman et al., 2020). It is a stochastic decoding scheme and is different from Top-k sampling, which relies on selecting a fixed number of tokens (top k) as samples according to their relative probabilities at each time-step. With nucleus sampling, for a given probability distribution conditioned on the previous words and the context, the top-p vocabulary is the smallest set $V^{(p)} \subset V$ that satisfies Equation 4.3, where x is the next word and p is the minimum probability. Figure 4.2 depicts an example of two time-steps in the nucleus sampling method and Figure 4.3 shows a cherry-picked example of generated text, based on different decoding algorithms.

$$\sum_{x \epsilon V^{(\mathrm{p})}} P(x|x_{1 \ldots t-1}) \geq p \tag{4.3}$$

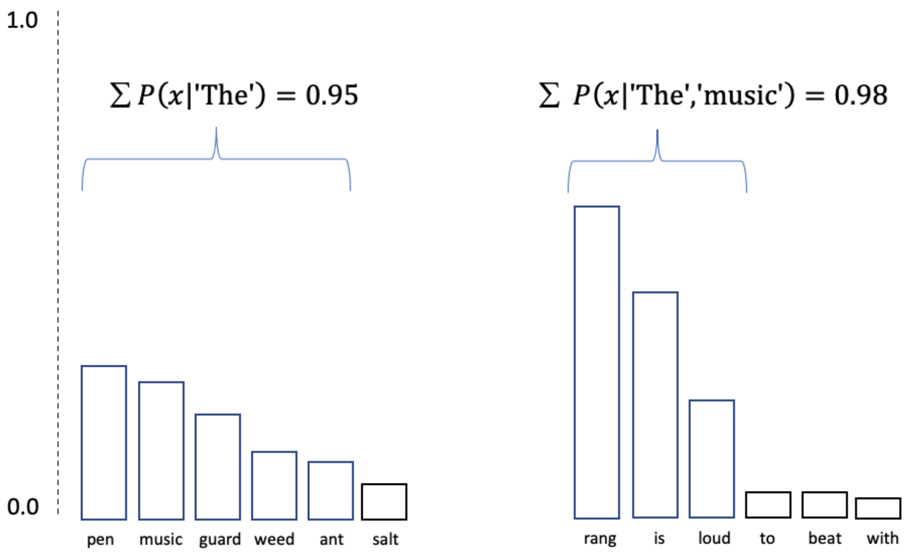

Figure 4.2: Nucleus (Top-p) sampling example for $p = 0.93$

In addition to the various decoding algorithms for generation, there are other important factors to consider for response generation. Temperature is one of them. This tilts the distribution towards highly probable samples, thereby lowering the mass in the tail distribution and controlling the shape of the distribution (Holtzman et al., 2020).

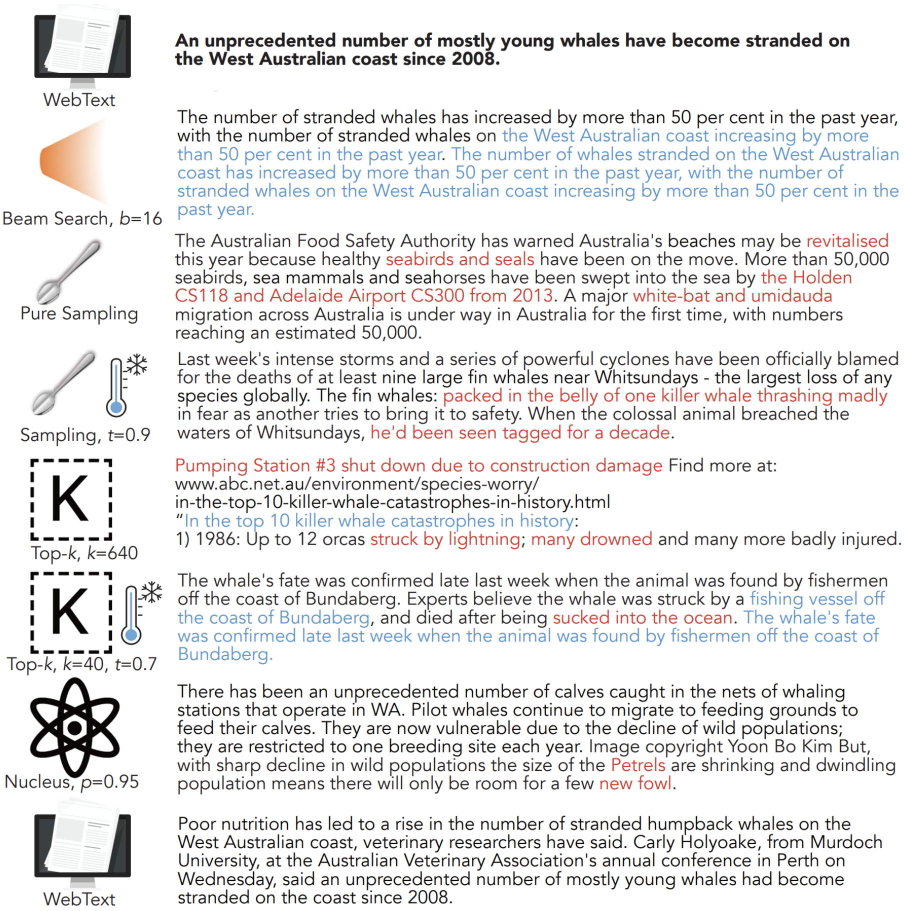

Figure 4.3: Cherry-picked example of comparison of decoding algorithms when a webtext context is provided. Red highlights show incoherence while blue highlights show unnecessary repetitions. Image from Holtzman et al. (2020)

4.3 Deep models for open-domain conversational systems

An NN is an adaptive and fairly complex system, as described in Section 1.6. Deep learning uses statistical techniques, based on sample data, for classifying patterns or making predictions by using NN with multiple layers. For these networks to generalise well, there must be large enough data, usually, and the test data should be similar to the training data, so that appropriate interpolation can be achieved (Marcus, 2018). Models based on reinforcement learning (RL) or adversarial networks are also used in the development of conversational systems (Adiwardana et al., 2020; Chowdhary, 2020; Jurafsky and Martin, 2020), however, our attention here will be on common models based on the encoder-decoder architecture or one of its stacks, usually the decoder. RL systems use rewards that are given at the end of a successful conversation, to train a policy to take action. Noteworthy that challenges still exist generally with deep learning models (Marcus, 2018) and some of them include struggling with open-ended inference, being data-intensive, requiring so many parameters that may impede transparency, engineering difficulty, and lack of commonsense reasoning (Bird et al., 2009). Below are some deep model architectures for open-domain conversational systems.

4.3.1 Encoder-Decoder

The encoder-decoder architecture conditions on the encoding of the queries and responses up to the last moment in order to generate the next response token (Jurafsky and Martin, 2020). It is common for generating conversations or responses to utterance prompts and is a sequence-to-sequence (seq2seq) model (Holtzman et al., 2020). A seq2seq model makes predictions by outputting a probability distribution over possible next response tokens (Adiwardana et al., 2020). The basic architecture is known for dull, repetitive responses (Chowdhary, 2020). IR techniques, like concatenation of retrieved sentences from Wikipedia to the dialogue context, is one way of augmenting the architecture for refined responses (Jurafsky and Martin, 2020).

Other shortcomings may be addressed by switching the objective function to a mutual information objective or introducing the beam search decoding algorithm to achieve relatively more diverse responses (Chowdhary, 2020). Both the encoder and decoder may use the LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Vaswani et al., 2017) as the base architecture. Some processes are basic to the encoder-decoder, regardless of the underlying architecture that is used. The sequence of words is run through an embedding layer in the encoder stack, which then compresses the sequence in the dense feature layer into fixed-length feature vector. The decoder produces a sequence of tokens after they are passed from the encoder layer. This is then normalised using a Softmax function, such that the word with the highest probability becomes the output. Attention (Bahdanau et al., 2015) may be introduced to the model. The attention mechanism focuses on desired parts of a sequence regardless of where they may appear in the input and ignores other parts or assigns less weighted average to them (Raffel et al., 2020).

4.3.2 DLGNet

DLGNet was presented by Olabiyi and Mueller (2019). It has a similar architecture as GPT-2. It is a multi-turn dialogue response generator that was evaluated, using BLEU, ROUGE, and distinct n-gram, on the Movie Triples and closed-domain Ubuntu Dialogue datasets. As an autoregressive model, it uses multiple layers of self-attention to map input sequences to output sequences by shifting the input sequence token one position to the right so that, at inference, the model uses the previously generated token as additional input for the next token generation. Instead of modelling the conditional distribution of the response, given a context, it models the joint distribution of the context and response. Two sizes of the model were trained: a 117M-parameter model and the 345M-parameter model, with 12 attention layers and 24 attention layers, respectively. No preprocessing of the datasets was done because of the use of BPE, which provided 100% coverage for Unicode texts and prevented the OOV problem. The good performance of the model is due, in addition to BPE, the long-range transformer architecture and the injection of random informative paddings.

4.3.3 Meena

Adiwardana et al. (2020) presented Meena, a multi-turn open-domain conversational agent that was trained end-to-end, being a seq2seq model (Bahdanau et al., 2015). The underlying architecture of the seq2seq model is the Evolved Transformer (ET). It has 2.6B parameters including 1 ET encoder stack and 13 ET decoder stacks. The hyper-parameters of the best Meena model were decided through manual coordinate-descent search. The data it was trained on is a filtered public domain corpus of social media conversations containing 40B tokens. Besides automatic evaluation, using perplexity, it was also evaluated in multi-turn conversations using the human evaluation metric: Sensibleness and Specificity Average (SSA). This human evaluation combines two essential aspects of a human-like chatbot: making sense and being specific.

4.3.4 BlenderBot 2

Roller et al. (2020) pointed out the ingredients for their SoTA model BlenderBot, which comes in different variants. Some of the ingredients are empathy and personality, consistent persona, displaying knowledge, and engagingness. Three types of architecture, all based on the Transformer, were investigated: retrieval, generative, and a combination of the two, called retrieve-and-refine. The generative architecture is a seq2seq model and uses Byte-Level BPE for tokenisation. Three variants, based on different number of parameters, were designed: 90M, 2.7B, and 9.4B. Human evaluation of multi-turn conversations, using ACUTE-Eval method, showed its best model outperformed previous SoTA on engagingness and humanness. The other main conclusions from their study are that finetuning on data that emphasises desired conversational skills brings improvement and models may give different results when different decoding algorithms are used, though the models may report the same perplexity.

4.3.5 Text-to-Text Transfer Transformer (T5)

Among the models that are pretrained on large text and may be adapted for conversational systems is Text-to-Text Transfer Transformer (T5) by Raffel et al. (2020). It is an encoder-decoder Transformer like the one by Vaswani et al. (2017) and depicted in Figure 3.5. An input sequence is mapped to a sequence of embeddings in the encoder, which is then fed to the decoder before the final dense Softmax layer. A simplified version of layer normalisation is employed such that no additive bias is used. The self-attention of the decoder is a form of autoregressive or causal self-attention. All the tasks considered for the model are cast into a text-to-text format, in terms of input and output. Maximum likelihood is the training objective for all the tasks but a task-prefix is specified in the input before feeding the model in order to identify the task at hand. The base version of the model has about 220M parameters.

4.3.6 GPT-3

Brown et al. (2020) introduced GPT-3, being the biggest size out of the eight models they created. It is an autoregressive model with 175B parameters that shares many of the qualities of the GPT-2. These include modified initialisation, pre-normalisation, and reversible tokenisation. It, however, uses alternating dense and locally banded sparse attention. Results in few-shot inference reveal that the model achieves strong performance on many tasks. Zero-shot transfer involves providing text description of the task to be done, during evaluation. This is different from one-shot or few-shot transfer, which involves conditioning on 1 or k number of examples for the model in the form of context and completion. No weights are updated in any of the three cases at inference time and there's a major reduction of task-specific data that may be needed. Despite the successes of the model, it struggles at few-shot learning with some datasets, loses coherence over sufficiently long passages, gives contradictory utterances, and its size makes it difficult to deploy.

4.3.7 DialoGPT

Dialogue Generative Pre-trained Transformer (DialoGPT) was trained on Reddit conversations of 147M exchanges (Zhang et al., 2020). It is an autoregressive LM based on GPT-2, another SoTA model (Radford et al., 2019). In single-turn conversations, it achieved performance close to human in open-domain dialogues, besides achieving SoTA in automatic and human evaluation. The medium model has 345M parameters and 24 transformer layers while the small model has 12 layers. In the model, a multiturn dialogue session is framed as a long text and the generation as language modelling. Furthermore, it employs what is called maximum mutual information (MMI) scoring to address the problem of bland response. This technique uses a pretrained backward model to the source sentences from the responses. An advantage of the model is the easy adaptability to new dialogue datasets with few samples. The more recent improvements to the DialoGPT model jointly trains a grounded generator and document retriever (Zhang et al., 2021). This is the predominant model that is used in the conversational systems experiments of this thesis. Figure 4.4 shows some of the hyperparameters set for the model in the experiments. The *no_repeat_ngram_size* determines the minimum length of the n-gram that should occur only once in the generated output.

```
model.generate(
    bot_input_ids,
    max_length=200,
    pad_token_id=tokenizer.eos_token_id,
    no_repeat_ngram_size=3,
    do_sample=True,
    top_k=100,
    top_p=0.7,
    temperature = 0.8
    )
```

Figure 4.4: Some hyperparameters for DialoGPT in this work.

4.3.8 Model cards

Model cards are the documentation or statements which detail the performance characteristics of ML models, according to Mitchell et al. (2019). They are necessary for these models because of the implications or outcome of using the models. They are useful for transparency. Model cards should not serve as disclaimer or exoneration from responsibility for strongly harmful or unethical models. They should provide evaluation information of the different conditions that may be applicable to the model. The context of use for the model, performance evaluation procedure, used metrics, and types of possible errors are also important in the model card. The importance of model cards, like their data counterpart mentioned in Section 2.6, cannot be over-emphasised. The discovery of systematic biases, such as those in face detection or criminal justice, have made this even more important (Mitchell et al., 2019). This is the reason some have called for algorithmic impact statements (Bender and Friedman, 2018). Mitchell et al. (2019) recommended the following additional details in a given model card under relevant sections: the person or group behind the developed model, versioning, licence, fairness constraints, intended use and users, demographics, training and evaluation data, ethical consideration, and recommendations. Not all sections of the model card may be relevant for every model. The appendix contains the model cards of some of the models used in this work. In particular, Appendix D for , E for , F for and G for .

4.4 Measuring progress

We need to measure the performance of any system to determine how successful it is. Since the goals of task-based systems are different from those of open-domain conversational systems, they do not always use the same evaluation metrics. Automatic evaluation metrics used in NLG tasks, like MT, such as BLEU or ROUGE, are sometimes used to evaluate conversational systems (Zhang et al., 2020). However, these metrics are also discouraged because they do not correlate well with human judgment (Jurafsky and Martin, 2020). Perplexity is sometimes used and has been shown to correlate with the human evaluation metric SSA (Adiwardana et al., 2020). Equation 4.4 is the mathematical equation of perplexity. It measures how well a model predicts the data of the test set, providing an estimate on how accurately it expects the words people will say next (Adiwardana et al., 2020). Very low perplexity for generated text, however, has been shown to imply such text may have low diversity and unnecessary repetition (Holtzman et al., 2020).

$$PP(W_{test}) = \sqrt[N]{\left(\frac{1}{\rho(W_{test})}\right)} \tag{4.4}$$

The most credible way, perhaps, for evaluating open-domain conversational systems (or chatbots) is through human evaluation. This may be done through participatory or observer evaluation. The participatory approach requires an evaluator to have a chat or conversation with the system while the observer approach requires a third party to read a transcript of conversations (Jurafsky and Martin, 2020). Some of the qualities that open-domain conversational systems may be evaluated on include: humanness (or human-likeness), engagingness, fluency, making sense, interestingness, and avoiding repetition. The Likert scale is usually provided for grading these various qualities. Most of the human evaluation in this work are based on human-likeness. The others are comparison of diversity and how fitting responses are to the given contexts. In some brief detail,

- human-likeness attempts to determine if the turns or conversations are the way humans would generally speak.
- engagingness attempts to establish if the conversation is engaging such that subsequent turns elicit continual user response so that the conversation lasts for a reasonable amount of time.
- fluency measures how fluent or articulate the generated turns or conversations are.
- making sense attempts to establish if the responses or the conversation is logical.
- interestingness may be considered closely related to engagingness and it attempts to determine if the turns or conversations are of interest.
- avoiding repetition evaluates if the generated text has unnecessarily repeated tokens.

4.5 Metaphors in the mouths of chatbots

It has been shown that metaphors have more emotional impact than their literal equivalent (Mohammad et al., 2016). Idioms generally make utterances more colorful (or rich) and diverse. Indeed, Holtzman et al. (2020) observed that the distribution of generated text (from beam search or pure sampling) is different and less surprising than natural text. In this work, as results in Section 4.6 reveal, the use of idioms appears to enrich and bring diversity to generated text, without changes to the decoding algorithm.

Jhamtani et al. (2021) asserted that robust performance of dialogue systems is dependent on the ability to handle figurative language. In order to use the PIE-English idioms corpus for training as intended in this work, we make the assumption that the corpus is suitable as a conversational dataset of dialogue turns, though the corpus is not originally a dialogue dataset. This assumption is valid because the sentences of the turns discuss the same cases of idioms despite being drawn from different examples from the base corpora: the BNC and UKWaC (Ferraresi et al., 2008).

4.6 Experiments & Evaluation

The experiments were set up to test the first hypothesis in Section 1.5. We utilise the model checkpoint by Adewumi et al. (2022c), which is already trained on the MultiWOZ and available on the HuggingFace hub Wolf et al. (2020) to produce another model checkpoint (IdiomWOZ) by finetuning on the PIE-English idioms corpus. A second new model checkpoint is created (IdiomOnly) from the original DialoGPT model by Zhang et al. (2020) by finetuning also on the same idioms corpus. The DialoGPT model for the second model checkpoint is the same medium variant utilised by Adewumi et al. (2022c) to produce the MultiWOZ checkpoint. The idioms corpus was split in the ratio 80:10:10 for the training, dev, and test sets, respectively, and multiple runs (3) per experiment conducted in order to determine the average perplexities and standard deviation.

The two newly saved model checkpoints from each category plus the MultiWOZ model checkpoint from Adewumi et al. (2022c) are then used to generate three conversation transcripts in a first set of experiments. Ninety-four random numbers were generated and used to select the same prompts from the test sets (the PIE-English idioms corpus and the MultiWOZ) to feed the three models. Thirty-two prompts for generation and fifteen prompts with their test set responses (for credibility) are selected from each test set. In the second set of experiments, sixty-two random numbers were generated. Thirty-two (from the idioms corpus) were used as prompts for two of the models (IdiomWOZ and MultiWOZ) while thirty are credibility conversations from the MultiWOZ test set.

The credibility conversations are to test the evaluators for their competence, hence the responses to these prompts are not generated but are the responses from the corresponding test sets. They are distributed at regular intervals within each transcript. All the experiments were run on a shared DGX-1 machine with 8 x 32 Nvidia V100 GPUs. The operating system of the machine is Ubuntu 18 and it has 80 CPU cores. From Table 4.1, which compares the average perplexity of the models, we observe that the MultiWOZ model from Adewumi et al. (2022c) has the lowest perplexity. This is very likely because the MultiWOZ data the model was trained on is larger (with more conversation turns) than the idioms corpus. The results are statistically significant as the p-value ($p < 0.0001$) of the two-sample t-test for the difference of two means (for the IdiomWOZ and IdiomOnly) is smaller than alpha (0.05). Although the average perplexity for the IdiomOnly model is lower than the IdiomWOZ, we chose to generate responses and have human evaluation on the latter, especially as one of its runs had a lower perplexity, as may be deduced from the standard deviation. In addition, perplexity alone does not tell how good a model is Roller et al. (2021); Hashimoto et al. (2019).

Table 4.1: Average perplexity results. sd - standard deviation

Model	**Perplexity**	
	dev (sd)	test (sd)
IdiomWOZ	201.10 (34.82)	200.68 (34.83)
IdiomOnly	189.92 (1.83)	185.62 (2.05)
MultiWOZ (Adewumi et al., 2022c)	6.41 (-)	6.21 (-)

Tables 4.2 and 4.3 present human evaluation results for two different transcripts of 64 and 32 single-turn conversations for the first and second set of experiments, respectively, after removing the 30 credibility conversations from each. Instruction 1 and Instruction 2 below are the instructions for the first and second set of transcripts, respectively. As Alm-Arvius (2003) speaks of the diverse types of meaningful variation in text, we evaluate the second transcript (with results in Table 4.3) based on two characteristics: more fitting and more diverse responses. Table 4.2 is based on humanlikeness. We observe that, under majority votes, two (MultiWOZ and IdiomWOZ) out of three of the models have more humanlike single-turn conversations than other categories. The MultiWOZ model has the most humanlike single-turn conversations. However, when we consider idioms only prompts in Table 4.2, the IdiomWOZ model has the most humanlike conversations. In Table 4.3, IdiomWOZ has more fitting conversations than the MultiWOZ, though the converse is the case with regards to more diverse conversations. This may be due to the evaluators' understanding or interpretation of what is diverse. For all the evaluations, we observe that there is CUS of 80%. The CUS is the same across sections in each table since the same transcript is involved for each section. Tables 4.4 and 4.5 show some single-turn conversations from the second transcript. Person 1 is the prompt from the PIE-English idioms test set.

> ***Instruction 1:*** Here are 94 different conversations by 2 speakers. Please, write Human-like (H) or Non-human-like (N) or Uncertain (U), based on your own understanding of what is human-like. Sometimes the speakers use idioms. If you wish, you may use a dictionary.

> ***Instruction 2:*** Person 2 & Person 3 respond to Person 1. Please, write which (2 or 3) is the a) more fitting response & b) more diverse response (showing variety in language use).

Table 4.2: Human evaluation results of 3 annotators on 3 classes for 64 single-turn conversations.

Model	**Scale (majority votes)**				**CUS**
	H (%)	U (%)	N (%)	3-way (%)	%
IdiomWOZ	39.1	10.9	37.5	12.5	80
IdiomOnly	15.6	12.5	60.9	10.9	80
MultiWOZ	62.5	1.6	32.8	3.1	80
unanimous votes					
IdiomWOZ	20.3	0	12.5	-	80
IdiomOnly	6.3	0	31.3	-	80
MultiWOZ	45.3	0	23.4	-	80
idioms only maj. votes (32 samples)					
IdiomWOZ	30	23.3	33.3	13.3	80
IdiomOnly	26.7	0.2	36.7	16.7	80
MultiWOZ	26.7	3.3	66.7	3.3	80

Table 4.3: Human evaluation results of 3 annotators on 3 classes for 32 single-turn conversations.

Model	**Scale (majority voting)**		**CUS**
	More fitting (%)	More diverse (%)	%
IdiomWOZ	71.9	28.1	80
MultiWOZ	28.1	71.9	80
unanimous votes			
IdiomWOZ	46.9	9.4	80
MultiWOZ	9.4	15.6	80

4.6.1 Evaluator feedback

Feedback from some of the evaluators suggests the use of idioms complicated the task a bit but it was sometimes influential in spotting which of the two conversations was a more fitting or diverse option. Evaluators found it relatively easier evaluting the humanlikeness characteristic in the first set of experiments but had to resort to a dictionary for the idioms in some instances.

4.7 Ethics of developing conversational systems

It has been mentioned earlier that ethical issues are an important consideration when developing conversational systems. In Section 5.4 we show examples of cases when models may be prompted with offensive text. Some systems have been known to demean or abuse their users. It is also well known that machine learning systems reflect the biases and toxic content of the data they are trained on (Neff and Nagy, 2016; Jurafsky and Martin, 2020). Privacy is another crucial ethical issue. Data containing PII may fall into

Table 4.4: Excerpt from the second transcript of single-turn conversations. Person 2 is the IdiomWOZ

Person 1	**Person 2**
He knew his father's grievance ; and put off the evil day as long as he possibly could	he's dead meat
achieving results that fly in the face of realism	I'm not here I'm just a fly on the wall
does that mean Jane is off the hook ?	She had the nerve to call time on her three-year spell in charge
he's still sulking ; sounding like a broken record	the whole caboodle
we need new ways to do this ; to chew the fat ; to check ourselves out	the Government's approach to the environment is geared more towards addressing the environment rather than helping the environment
the 24-year-old striker is self-critical enough to insist he is not yet firing on all cylinders	You're the one up to your neck in it
remind her to take it off the hook	the bookie took it on the nod
A lengthy break is in order ; so make sure you keep your nose clean	they are under a cloud because of fears about US President's wife Hillary Clinton
all the young men wanted to be in on it before it was all over	It was the last word in a book he had bought for a song
The phrase ; 'There's one born every minute ' comes to mind .	I'm going to jump the gun and name my 12th man
Some families breed like rabbits and send us all their male heirs	he was out for the count
Utd have shown they are in a league of their own	It's a long shot him signing for Leeds. They're a hard nut to crack
from now until the year dot	he was the apple of her eye
The founders of Press for Change ended the tyranny of "fear of exposure" by showing it for what it was – a paper tiger	She felt hot and heavy under the hot weather
it's a good idea to go the distance	it 'll just collapse and give up the ghost somewhere in South Africa

the wrong hands and cause security threat to those concerned. It is important to have systems designed such that they are robust to such unsafe or harmful attacks. Attempts are being made with debiasing techniques to address some of these challenges (Dinan et al., 2020). Privacy concerns are also being addressed through anonymisation techniques (Henderson et al., 2018; Jurafsky and Martin, 2020). Balancing the features of chatbots with ethical considerations can be a delicate and challenging work. For example, there is

Table 4.5: Excerpt from the second transcript of single-turn conversations. Person 3 is the MultiWOZ

Person 1	**Person 3**
He knew his father's grievance ; and put off the evil day as long as he possibly could	That's great; thanks.
achieving results that fly in the face of realism	Yes; realism. I like it.
does that mean Jane is off the hook ?	I'm looking for a place to go in the centre.
he's still sulking ; sounding like a broken record	Well then I suppose I need to find something else to do for you.
we need new ways to do this ; to chew the fat ; to check ourselves out	To each their own. Let me know if you need any more info.
the 24-year-old striker is self-critical enough to insist he is not yet firing on all cylinders	Well if you'll excuse me... let me ask my wife. Does she think we should try for a little earlier; say 20:30?
remind her to take it off the hook	I'm sorry; I forgot to hang up. Good bye.
A lengthy break is in order ; so make sure you keep your nose clean	I am looking for a restaurant in the centre of town that serves unusual food.
all the young men wanted to be in on it before it was all over	I was able to book the table for you. The reference number is : RLTOT7X2. Can I help with anything else?
The phrase ; 'There's one born every minute ' comes to mind .	I was able to book the table for you. The reference number is : RLTOT7X2. Can I help with anything else?
Some families breed like rabbits and send us all their male heirs	I'm sorry; but I'm not sure what you're looking for. Could you please restate your requirements?
Utd have shown they are in a league of their own	I'm sorry; but I'm not sure what you're looking for. Could you please restate your requirements?
from now until the year dot	I'm sorry; but I don't understand your request. Would you like to try a different restaurant?
The founders of Press for Change ended the tyranny of "fear of exposure" by showing it for what it was – a paper tiger	Excellent. Can I have the address and postcode; please?
it's a good idea to go the distance	I can help with that! What are your departure and destination sites?

contention in some quarters whether using female voices in some technologies/devices is appropriate. Then again, one may wonder if there is anything harmful about that. This is because it seems to be widely accepted that the proportion of chatbots designed as "female" is larger than the those designed as "male". In a survey of 1,375 chatbots, from automatically crawling chatbots.org, Maedche (2020) found that most were female.

We surveyed 100 chatbots[4], with regards to the binary gender: female or male, across different literature. The method employed for the survey involved the initial step of searching based on the term "gender chatbot" in Google scholar and recording details of all chatbots mentioned in the first 10 pages of the search result. Thereafter, the Scopus database was equally queried with the same search term and it turned out 20 links. Both sites resulted in 120 result links from which 59 chatbots were identified. Facebook Messenger, which is linked to the largest social media platform, was chosen to provide additional 20 chatbots. The chatbots from Facebook Messenger were selected from 2 websites that provided information on some of the best chatbots on the platform[5]. The sites were identified with the search term "Facebook Messenger best chatbots" on Google and the chatbots were selected based on the first to appear on the list. Meanwhile, 13 chatbots have won the Loebner prize in the past 20 years, as some are repeat winners. Some chatbots mentioned in the scientific literature hosted their chatbots on Facebook Messenger but are not counted twice in this survey. This is also true for Loebner prize chatbots mentioned in the scientific papers. The 8 popular/commercial chatbots in the survey include Microsoft's Cortana and XiaoIce, Apple's Siri, Amazon's Alexa, Google Assistant, Watson Assistant, Ella, and Ethan by Accenture.

Each chatbot's gender is identified by the designation given by the developer or cues such as avatar, bot name or voice, especially in cases where the developer did not specifically identify the gender of the chatbot. These cues are created based on general perception or stereotypes. A chatbot is considered genderless if it is specifically stated by the reference or developer or nothing is mentioned about it and there are no cues to suggest gender. Maedche (2020) uses similar cues in their research. Technically, creating gendered chatbots through ML involves training computer models with data attributed to a particular gender, such as using samples of female voice to train a chatbot to have female voice. Overall, in our survey of the 100 chatbots, 37 (or 37%) are female, 20 are male, 40 are genderless, and 3 have both gender options. When the data is further broken down into 4 groups: journal-based, Loebner-winners, Facebook Messenger-based, and popular/commercial chatbots, we observe that one constant trend is that female chatbots always outnumber male chatbots. Even the genderless category does not follow such a consistent trend in the groups. Out of the 59 chatbots mentioned in journal articles, 34% are female, 22% are male, 42% are genderless, and 2% have both gender options. 54% are female among the 13 chatbots in the Loebner-winners, 23% are male, 15% are genderless, and 8% have both options. Of the 20 chatbots from Facebook Messenger, 25% are female, 10% are male, 65% are genderless, and 0 offer both genders. Lastly, of the 8 popular/commercial chatbots, 62.5% are female, 25% are male, 0 is genderless, and 12.5% have both options.

The results support the popular assessment that female chatbots are more predomi-

[4]May, 2020.

[5]growthrocks.com/blog/7-messenger-chatbots
enterprisebotmanager.com/chatbot-examples

nant than the male chatbots. Although we do not have information on the gender of the producers of these 100 chatbots, it may be a safe assumption that most are male. This observation of the predominance of chatbots being female has faced criticism in some quarters, such as a recent report by West et al. (2019) that most chatbots being female makes them the face of glitches resulting from the limitations of AI systems. Despite the criticism, there's the argument that this phenomenon can be viewed from a vantage position for women, such as being the acceptable face, persona or voice, as the case may be, of the planet. Silvervarg et al. (2012) compared a visually androgynous agent with both male and female ones and found that it suffered verbal abuse less than its female counterpart but more than the male. Does this suggest developers should do away with female chatbots altogether to protect them or what we need is a change in the attitude of users? This is especially given that previous research has shown that stereotypical agents, with regards to task, are often preferred by users (Forlizzi et al., 2007). Some researchers have argued that chatbots having human-like characteristics, including gender, builds trust for users (Louwerse et al., 2005; Muir, 1987; Nass and Brave, 2005). Also, Lee et al. (2019) in their study, observed that chatbots that consider gender of users, among other cues, are potentially helpful for self-compassion of users. An interesting piece of research might be to give consumers the option to choose chatbot gender, find out what the total distribution will be and ascertain the reasons for users' choices. It should be noted that there are those who find the ungendered, robotic voice of AI eerie and uncomfortable and will, thus, prefer a specific gender.

Chapter 5

Learning Deep Abstractions

"Models are like the brain."

(Simile)

While working on cross-lingual transferability, Artetxe et al. (2020) hypothesised that deep monolingual models learn some abstractions that generalise across languages. This may contrast with the previous hypothesis that attributes the generalisation capability of deep multilingual models to the shared subword vocabulary that is used across the languages, and their joint training, as demonstrated for mBERT (Pires et al., 2019). The performance of these models on low-resource languages and unseen languages are known to be relatively dismal, especially when compared to their monolingual counterparts (Pfeiffer et al., 2020; Wang et al., 2021; Virtanen et al., 2019; Rönnqvist et al., 2019). Furthermore, the multilingual versions of the deep models do not cover all languages, meaning many languages are still under-represented.

In this chapter, we will explore the commonalities in human languages first before looking at pretraining for transfer learning in Section 5.2 and multilingual deep models in Section 5.3. Thereafter, results from the experiments and evaluation on cross-lingual transferability are presented in Section 5.4.

5.1 Commonalities in human languages

Language may be described as the use of a finite set of elements (e.g. words), and making a set of rules (grammar and syntax) to create different comprehensible combinations for communication[1]. It is the principal mode of human communication, according to Google/Oxford Languages, consisting of words that are used in a conventional and structured way and conveyed by writing, speech or gesture[2] (Friederici, 2017). Although there are over 6,000 languages in the world with their peculiarities (Futrell et al., 2015; Youn et al., 2016), there is strong evidence that suggests many of them share certain

[1]bbc.com/future/article/20121016-is-language-unique-to-humans

[2]Google/Oxford Languages, accessed on April 6, 2022.

common features. Friederici (2017) believes that similarities in the structure which many languages share may be a result of how quickly and accurately the brain likes to process information. She refers to this underlying commonalities as "linguistic universals" or "cross-linguistic generalisations". Fitch (2011) calls them "formal universals" and thinks they may be understood as the model of a general solution to a set of differential equations, where each language is one particular solution. Two pointers to these linguistic universals are semantic similarity across languages through polysemous words (Youn et al., 2016) and minimal dependency length (MDL) (Futrell et al., 2015).

Youn et al. (2016) provide an empirical measure of semantic proximity among concepts by using crosslinguistic dictionaries for translation of words between languages. It involves observation of polysemies (words having more than one meaning) in the vocabulary across different language groups, which shows that the structural properties are consistent across the language groups, and largely independent of environment. The frequency of two concepts sharing a single polysemous word in a sample of unrelated languages determines the measure of semantic similarity between them (Youn et al., 2016). The study focused on a sample of 81 languages in a phylogenetically and geographically stratified way. The 81 languages include the Hausa, Yorùbá, and Swahili languages, which are examined in this thesis, where Yoruba and Swahili are grouped under the Niger-Kordofanian family and Hausa is in the Afro-Asiatic family (Youn et al., 2016). They noted that a group of languages may have structural resemblances as a result of the different speakers having common historical or environmental features. Figure 5.1 shows part of the universal semantic network of languages.

For quantitative, cross-linguistic evidence of MDL, Futrell et al. (2015) provide a relatively large-scale demonstration for this syntactic property of languages, showing that dependency lengths are shorter than chance. MDL is the attempt to reduce the distance between syntactically related words in a sentence (Futrell et al., 2015). Distances between linguistic heads and their dependents in a sentence are called dependency lengths, where the head licenses another word (the dependent). It supports previously held view that speakers prefer short dependency length in word orders and that languages tend to follow the same direction. In the study, which involves 37 languages, including English and Swedish, which are part of the investigation in this thesis, it is shown that the overall dependency lengths are shorter than random baselines by conservative estimates, for all the languages. This suggests that MDL is a universal quantitative property of human languages. It is a functional explanation that the grammars of languages evolved in order that users of languages may communicate through sentences that are relatively easy to produce and understand.

MDL is seen as a reliable generalisation in NLP, as observed by Futrell et al. (2015), since many SoTA models incorporate a bias in favour of positing short dependencies (Klein and Manning, 2004; Smith and Eisner, 2006). This chapter evaluates cross-lingual transferability from English for seven target languages, possibly exploiting these linguistic universals,. These target languages are Swedish, Swahili, Wolof, Hausa, Kinyarwanda, Yorùbá, and Nigerian Pidgin English. The languages are briefly discussed in the following subsections. The target languages cover Sweden and Finland, shown in Figure 5.2, and

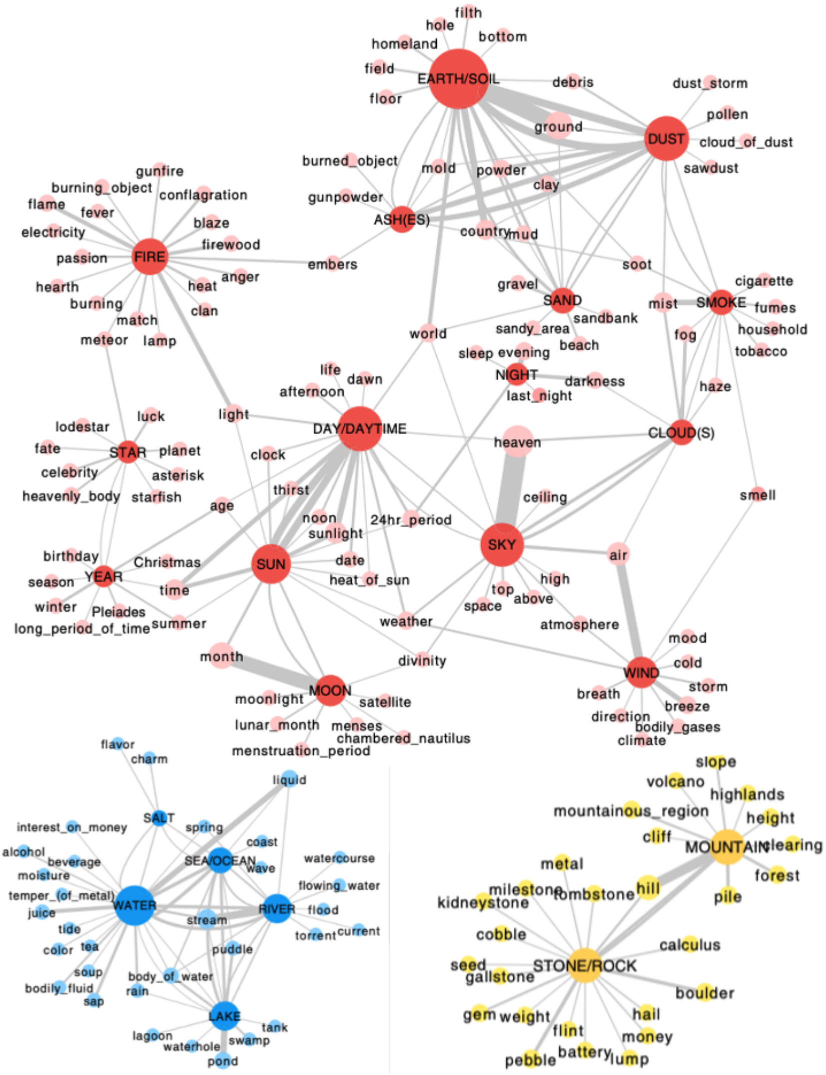

Figure 5.1: The universal semantic network of languages, based on polysemy (Youn et al., 2016). Concepts are linked when polysemous words cover both concepts. Swadesh words (the starting concepts) are capitalized. The size of a node and the width of a link to another node are proportional to the number of polysemies associated with the concept and with the two connected concepts, respectively. This distribution indicates that concepts have different tendencies of being polysemous. For example, EARTH/SOIL has more than 100 polysemies, whereas SALT has only a few. Three distinct clusters, colored in red, blue, and yellow, are identified.

countries in West, East, Central, and Southern Africa, shown in Figure 5.3 (Heine et al., 2000). The target languages involve a total of over 249 million speakers.

Figure 5.2: Sweden and Finland. Image from online.seterra.com

5.1.1 English

Modern or standard English (subsequently referred to simply as English) is quite different from the English of the early periods (Crystal, 2018). It is one of the West Germanic languages belonging to the Indo-European language family[3]. Besides being the national or dominant language of England, Canada, and the United States of America, it is the lingua franca for many countries and many domains (Björkman, 2014). It is the world's most international language (Konig and Van der Auwera, 2013). Examples of English sentences from the MultiWOZ dataset are provided below.

[3] britannica.com/topic/English-language

- I have several options for you; do you prefer African, Asian, or British food?
- I want to book it for 2 people and 2 nights starting from Saturday.
- That is all I need to know. Thanks, good bye.

5.1.2 Swedish

The Swedish language is spoken by more than 8.5 million people in Sweden as a national language (Reuter, 1992). It is also one of the prominent languages of Finland (Konig and Van der Auwera, 2013). It is a Germanic language and bears resemblance with Danish and Norwegian for historical reasons (Konig and Van der Auwera, 2013). Below is the Swedish translation of the English sentences mentioned earlier, from the MultiWOZ dataset.

- Jag har flera alternativ för dig; föredrar du afrikansk, asiatisk eller brittisk mat?
- Jag vill boka det för 2 personer och 2 nätter från och med Lördag.
- Det är allt jag behöver veta. Tack hejdå.

5.1.3 Swahili

Swahili, a Bantu language, is predominant in the southern half of Africa (Polomé, 1967). It is also an official language for countries in the East African Community (EAC). The countries are Burundi, Uganda, South Sudan, Kenya, Tanzania, Rwanda, and the Democratic Republic of the Congo (DRC). Zambia, Mozambique, the southern tip of Somalia, and Malawi use the language as lingua franca (Polomé, 1967). Over 50 million people speak the language[4]. It is a working language of the African Union. Below is the Swahili translation of the English sentences mentioned earlier.

- Nina chaguzi kadhaa kwako; unapendelea chakula cha Kiafrika, Kiasia, au Uingereza?
- Nataka kuihifadhi kwa watu 2 na usiku 2 kuanzia Jumamosi.
- Hiyo ndiyo yote ninahitaji kujua. Asante, kwaheri.

5.1.4 Wolof

Wolof is used in Mauritania, Senegal, and the Gambia. It has more than 7 million speakers[5]. It is of the Senegambian branch of the Niger–Congo language phylum. It is the largest language phylum in the world (Heine et al., 2000). Wolof is not a tonal language, unlike most other languages of the Niger-Congo phylum. Below is the Wolof translation of the English sentences from the MultiWOZ dataset.

[4] swahililanguage.stanford.edu
[5] worlddata.info/languages/wolof.php

- amna ay tanneef yu bari ngir yaw. ndax bëg ngan lekku niit ñu ñull yi, wa asi wala wa angalteer
- Soxla jënd ngir ñaari niit ak ñaari guddi mu tambelee gawu
- dedet li rek la soxla. jerejef. ba benen yoon

5.1.5 Hausa

Hausa is spoken by the Hausa people and is a Chadic language, which is the most widely spoken language of the Chadic branch of the Afroasiatic phylum Heine et al. (2000). The northern part of Nigeria and the southern part of Niger are where it is mainly predominant but it has minorities in Cameroon, Benin, and Chad. There are more than 40 million speakers[6]. Below is the Hausa translation of the English sentences from the MultiWOZ dataset.

- Ina da zabubbuka da yawa a gare ku; kun fi son abincin Afirka, Asiya, ko Biritaniya?
- Ina so in yi wa mutane 2 da dare 2 farawa daga ranar Asabar.
- Wannan shine kawai abin da nake bukatar sani. Godiya, bye bye.

5.1.6 Nigerian Pidgin English

Nigerian Pidgin English is popular among young people and is a simplified means of communication among the ethnic groups in Nigeria. The vocabulary and grammar are limited and often drawn from the English language (Ihemere, 2006). About 75 million people are estimated to speak the language though the actual number is difficult to say[7]. Below is the Nigerian Pidgin translation of the English sentences mentioned earlier.

- I get plenty options for you! you prefer African, Asian, or British food?
- I wan book am for 2 people for 2 night for Saturday
- na everything wey i need to know. thank you. good bye

5.1.7 Kinyarwanda

Kinyarwanda is an official language of Rwanda. It is also a dialect of the Rwanda-Rundi language (Heine et al., 2000). More than 22 million people are estimated to be speakers of the language[8]. Below is the Kinyarwanda translation of the English sentences.

- Mfite henshi naguhitiramo hari ibiryo bitetse mu buryo bw' Afrika, Aziya, cyangwa Ubwongereza?

[6] britannica.com/topic/Hausa-language
[7] bbc.com/news/world-africa-38000387
[8] worlddata.info/languages/kinyarwanda.php

- Ndashaka kubika imyanya ku bantu 2 n'amajoro 2 guhera ku wa Gatandatu.
- Ibyo ni byo nari nkeneye kumenya. Urakoze, murabeho.

5.1.8 Yorùbá

Yorùbá is predominantly spoken in Southwestern Nigeria by the Yorùbá ethnic group (Heine et al., 2000). It is spoken in areas spanning Nigeria and Benin with smaller migrated communities in Sierra Leone, Cote d'Ivoire, and The Gambia. More than 45 million people are estimated to speak the language[9]. Below is the Yorùbá translation of the English sentences from the MultiWOZ dataset.

- Mo ní awọ́n àṣàyàn púpọ̀ fún ọ; ṣé o fẹ́ràn óunjẹ Áfríkà, Ásíà, tàbí ìlú Gẹ̀ẹ́sì?
- Mo fé ṣe ìwé fún ènìyàn méjì àti fún alẹ́ méjì tí ó bẹ́rẹ̀ láti ọjọ́ Sátìdeé.
- Ìyẹn ni gbogbo ohun tí mo nílò láti mọ̀. O ṣeun, Ó dàbò.

5.2 Pretraining for transfer learning

Erhan et al. (2010) observed that the best results in supervised learning tasks usually are brought about by an unsupervised learning component, which is an unsupervised pretraining phase. He et al. (2019), however, asserted that training from scratch (random initialisation) can often give similar performance as pretraining and finetuning, particularly in computer vision (CV). Others, like Hendrycks et al. (2019), disagree, showing that pretraining improves robustness. Even He et al. (2019) acknowledge that training from scratch will involve more number of training iterations (compared to finetuning) for the randomly initialized models to converge. The process of pretraining can be described by greedy layer-wise unsupervised training. Each layer learns a nonlinear transformation of its input, which is the output of the previous layer that captures the main changes in its input (Erhan et al., 2010). Some suggestions as to why pretraining works well are that 1) it is a conditioning or regularisation mechanism for the parameters of the network (Erhan et al., 2009, 2010) and 2) it is helpful for initialising the network around the parameter space where optimisation is easier, such that a better local optimum of the training criterion is found (Bengio et al., 2007).

There are several types of pretraining objectives (or tasks). Some of them include Masked Language Model (MLM) or denoising objective (Devlin et al., 2018a), Next Sentence Prediction (NSP) (Devlin et al., 2018a), Causal (or autoregressive) Language Model (CLM) (Brown et al., 2020; Zhang et al., 2020), Sentence Distance (Sun et al., 2020, 2021), Sentence Reordering (Sun et al., 2020, 2021), and Universal Knowledge-aware Pretraining (Sun et al., 2021). MLM randomly masks a small part of the input tokens, with the objective of predicting the original vocabulary id of the masked word

[9]worlddata.info/languages/yoruba.php

Figure 5.3: Coverage of the African languages in this thesis. Colors added only for aesthetics. Image from online.seterra.com

based only on its context. NSP determines if two sentences semantically follow each other or are related. Sentence Distance is an extension of NSP and is widely used in various pretrained models (Sun et al., 2021). Sentence Reordering learns relationship between sentences by reorganising permutated segments from a randomly split paragraph. Universal knowledge-aware pretraiing uses a pair of triples from knowledge graphs and the corresponding sentences from encyclopedia, where relation in triple or words are randomly masked. Pretraining of monolingual deep models for low-resource languages is a challenge because of the scarcity of data in such languages. This has motivated pretraining multilingual deep models.

5.3 Multilingual deep models

Multilingual deep models are deep models that are usually pretrained on unstructured data of two or more languages with the same pretraining task. Deep architectures are usually needed to learn the complicated functions that represent the high-level abstractions (Erhan et al., 2010). Some of these models are discussed briefly below and Table 5.1 summarises the languages represented in some multilingual models and Google MT.

Table 5.1: The languages in some models: √: yes, X: no (Adewumi et al., 2022a)

Language	Multilingual model					
	mBERT	mBART	mT5	XLM-R	AfriBERTa	Google MT
Swedish	√	X	√	√	X	√
Pidgin English	X	X	X	X	√	X
Yorùbá	√	X	√	X	√	√
Hausa	X	X	√	√	√	√
Wolof	X	X	X	X	X	X
Swahili	√	X	√	√	√	√
Kinyarwanda	X	X	X	X	X	√

5.3.1 Multilingual Text-to-Text Transfer Transformer (mT5)

Xue et al. (2021) introduced this multilingual variant of T5. It was pretrained on a large multilingual dataset (mC4) covering 101 languages. However, three of the languages in this thesis are not covered by mT5. These are Wolof, Nigerian Pidgin English, and Kinyarwanda. The pipeline follows the general-purpose text-to-text format and pretraining on unlabeled data without dropout. Data sampling for each language in the corpus employed a zero-sum strategy, thereby controlling the probability of training on low-resource languages to mitigate the possibility of overfitting for low-resource languages and underfitting for high-resource languages.

5.3.2 Multilingual Bidirectional Encoder Representations from Transformers (mBERT)

The multilingual version of BERT by Devlin et al. (2018a) is a pretrained model for 104 languages. It is trained on Wikipedia using the familiar MLM objective. BERT is an encoder stack from the Transformer architecture, where the large version has 24 stacks. It is pretrained with a deeply bidirectional method, where 15% of the words in the input is masked so that it predicts only the masked words. In mBERT, exponentially smoothed weighting of the data (and vocabulary creation) is performed. This is to balance the amount of data from high-resource and low-resource languages. High-resource languages will be under-sampled while low-resource languages will be over-sampled. For tokenisation, a 110K shared WordPiece vocabulary is used[10] and the same recipe as used for English is applied to all other languages so that 1) words are lower-cased and accents removed (though accent is important in some languages), 2) there's splitting of punctuation, and 3) tokenisation based on whitespace. The mBERT cased version fixes normalisation issues in a lot of the languages. Four of the languages in this work are not available in mBERT (Devlin et al., 2018b). They include Wolof, Hausa, Nigerian Pidgin English, and Kinyarwanda.

5.3.3 Multilingual Bidirectional & Auto-Regressive Transformer (mBART)

Liu et al. (2020b) presented mBART, a Transfromer-based seq2seq denoising auto-encoder, pretrained on monolingual corpora in 25 languages (Lewis et al., 2020). It is the first method for pretraining a seq2seq model by denoising full texts in several languages. It is trained once for all languages and provides a set of parameters that can be finetuned. Although mBART is pretrained on 25 languages from the common crawl corpora, none of the languages in the thesis are represented in mBART (Liu et al., 2020b).

5.3.4 Cross-Lingual Model-RoBERTa (XLM-R)

XLM-R is also a Transformer-based multilingual MLM that is pretrained on text from 100 languages (Conneau et al., 2020). The Common Crawl dataset used for training was more than two terabytes of filtered data but one dump was used for English while twelve dumps were used for all other languages. Subword tokenisation was directly applied on raw text data using SentencePiece. Language embeddings are not applied and it is assumed this allows the model to better deal with code-switching (the use of more than one language in one context). A vocabulary size of 250K was utilised. Conneau et al. (2020) observed that more languages in the multilingual model leads to better cross-lingual performance on low-resource languages up until a point. Again, four of the languages in this work are not available in XLM-R. They include Wolof, Yorùbá, Nigerian Pidgin English, and Kinyarwanda.

5.4 Experiments & Evaluation: Cross-lingual transferability

We demonstrate that generation of conversations is possible, with reasonable performance, for a foreign language though the pretraining was in English (Adewumi et al.,

[10] github.com/google-research/bert/blob/master/multilingual.md

2022c). This is done first for the Swedish language and then six other African languages, in a second set of experiments. The investigation seemingly agrees with the hypothesis that deep monolingual models learn abstractions that generalise across languages, as demonstrated also by Artetxe et al. (2020), though their experiments are different from those carried out in this thesis. Less computational effort was needed to demonstrate this hypothesis in this work. The models produced are hosted on the HuggingFace hub[11].

5.4.1 First experimental setup

DialoGPT (medium) model is used in the first set of experiments involving Swedish. Zhang et al. (2020) reported that the medium model gave the best performance when compared to its small and big variants. This is compared with a baseline seq2seq model that is trained on the Swedish GDC dataset. The seq2seq model is an LSTM architecture (Hochreiter and Schmidhuber, 1997) and uses the attention mechanism (Bahdanau et al., 2015), based on the ParlAI platform by Miller et al. (2017). It has 6M trainable parameters and a batch size of 64 is used to train it. The experiments were carried out on several Tesla V100 GPUs on an Nvidia DGX-1 machine running Ubuntu 18 and having 80 CPU cores.

The various Swedish datasets for the first set of experiments are conversational data from Reddit (2 sizes), Familjeliv (3 sizes) and the GDC (Allwood et al., 2003). These are shown in Table 5.2. They are later compared with the English MultiWOZ in perplexity results. The datasets are pre-processed by removing emails, URLs, numbers and some special characters. The datasets were split in the ratio 80:10:10 for training, dev, and test sets, respectively. The conversation context is 7 during training. Multiple runs (5) per experiment were conducted and the average perplexity reported in Table 5.3. As the the data size increases, the perplexity falls, as expected. Although the model trained on the MultiWOZ achieves the best perplexity, this is not unexpected, given that DialoGPT is pretrained on English data. The model trained on the Familjeliv size of over 1M turns is the best-performing of the Swedish models. The seq2seq model, whose architecture is different from DialoGPT and is not pretrained, has the worst perplexity result overall.

Table 5.2: Summary of datasets in first set of experiments (Adewumi et al., 2022c).

Dataset	File Size	Conversation Lines
Reddit 4K	0.57M	4,300
Reddit 60K	10.4M	59,437
Familjeliv 70K	10.3M	71,470
Familjeliv 400K	45.3M	347,590
Familjeliv 1M+	200M	1,576,360
GDC	6.6M	108,571
MultiWoZ (English)	11M	143,048

[11] huggingface.co/tosin

Table 5.3: Mean perplexity results for the different datasets after training for 3 epochs (Adewumi et al., 2022c)

Dataset	Dev set	Test set
Reddit 4K	71.94	88.31
Reddit 60K	65.86	51.70
Familjeliv 70K	11.12	12.27
Familjeliv 400K	7.02	7.44
Familjeliv 1M+	7.150	7.148
GDC	29.17	23.95
Seq2seq-GDC	2,864	2,865
MultiWOZ (English)	6.41	6.21

Table 5.4 shows results from human evaluation of the Swedish single-turn conversations for three of the models, with the last row showing scores for the original dialogue (i.e. human-human conversations from the Familjeliv test set). Single-turn conversations are evaluated, as practised in the original paper (Zhang et al., 2020), by drawing 30 prompts randomly from the test set for each model and generating responses from the model. A Likert scale of *clearly human-like (4.0), somewhat human-like (3.0), not very human-like (2.0), clearly not human (1.0)* was provided. The sum of columns 4.0 and 3.0 gives the human-likeness average for each model and the model trained on the familjeliv 1M+ dataset achieves 57.2% human-likeness score. Figures 5.4, 5.5, 5.6, 5.7, and 5.8 show conversations with the DialoGPT-MultiWOZ and the DialoGPT-GDC models. We observe from some of the conversations that when prompted with offensive phrases in both languages, the models do not respond back with offensive text. The response from the DialoGPT-GDC model is, however, less coherent than the DialoGPT-MultiWOZ. This is not surprising since the DialoGPT-MultiWOZ was trained on more data and performs better in perplexity result. The transcripts of the conversations are available online[12].

Table 5.4: Human evaluation score per model. clearly human-like (4.0), somewhat human-like (3.0), not very human-like (2.0), clearly not human (1.0) (Adewumi et al., 2022c)

Dataset	**Scale (%)**				
	1.0	2.0	3.0	4.0	sum(3.0,4.0)
Familjeliv 1M+	0.233	0.194	0.252	0.320	0.572
GDC	0.317	0.221	0.221	0.240	0.461
Reddit 60K	0.582	0.220	0.154	0.044	0.198
Human-human	0	0.079	0.281	0.640	0.921

[12]github.com/tosingithub/gemdesk

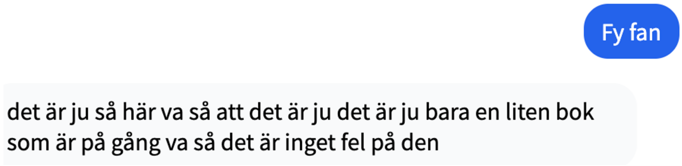

Figure 5.4: DialoGPT-GDC model prompted with a Swedish offensive phrase (Blue: user).

Figure 5.5: DialoGPT-GDC model prompted with certain Swedish phrases (Blue: user).

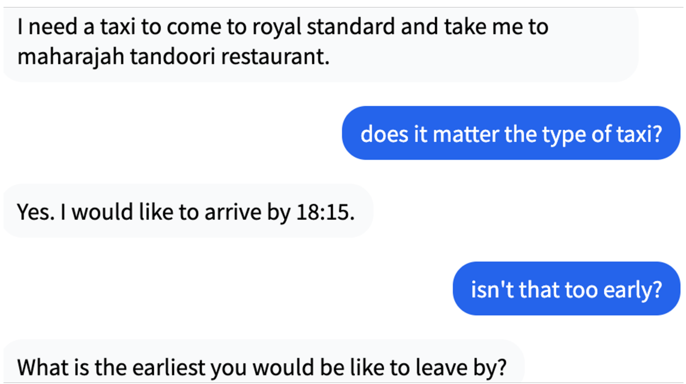

Figure 5.6: DialoGPT-MultiWOZ prompted with English sentences (Blue: user).

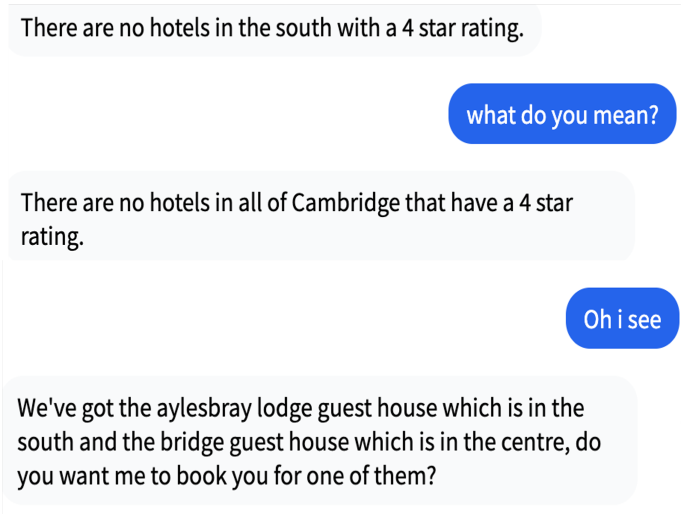

Figure 5.7: DialoGPT-MultiWOZ prompted with English sentences (Blue: user).

Figure 5.8: DialoGPT-MultiWOZ prompted with English offensive phrase (Blue: user).

5.4.2 Second experimental setup

In the second set of experiments, instead of two main models, an additional SoTA model is compared: BlenderBot 90M (Roller et al., 2021). The experiments were conducted using a participatory approach (Nekoto et al., 2020) on Google Colaboratory while some other experiments were run on the shared DGX-1 machine mentioned earlier. Each experiment was run 3 times and the average perplexity (including standard deviation) obtained. The training time for the BlenderBot 90M and the seq2seq models was for 20 minutes each. The decoding algorithm for all the models was set as top-k (k=100) and top-p (p=0.7).

The three models do not have exactly the same parameters or configuration and are not expected to have the same performance.

Method of human evaluation

Similar to the first set of experiments, we use the observer evaluation method, where evaluators read transcripts of conversations (Jurafsky and Martin, 2020). They rate single-turn conversations for human-likeness on a Likert scale with 3 entries (human-like (H), non-human-like (N) or uncertain (U)). A transcript is given to 3 native/L1 speakers per language to evaluate. Thirty-two single-turn conversations are generated per language and 3 credibility test dialogues spread out within the transcript to make up 35. A randomly generated list was used to select the same 32 prompts for all the languages from each test set of the AfriWOZ dataset. DialoGPT c7 x 1,000 (having context size 7 and 1,000 training turns), which had the best perplexity across languages, was used to generate the conversations, though small scale human evaluation is carried out to verify sample conversations from the other models: BlenderBot 90M and the seq2seq.

Eighteen conversation transcripts returned were credible out of the total of twenty-four. Discredited transcripts are the ones that failed 2 or more out of the 3 credibility test conversations by marking them as anything but H. The 3 credibility conversations are prompts and responses directly from the AfriWOZ test set instead of generated responses from the model. The evaluators were recruited on Slack[13]. They are also L1 speakers of the target languages and second/L2 (but dominant) speakers of English. They are not connected to the translation of the datasets nor did they take part in the training of the models, making them unbiased evaluators. The instruction for every evaluator at the top of the transcript of conversations is given below.

> Below are 35 different conversations by 2 speakers. Please mark each one as Human-like (H) or Non human-like (N) or Uncertain (U) based on your own understanding of what is human-like.

Table 5.5 gives the perplexity results for the three models. DialoGPT with a context size of 14 achieves the lowest perplexity per language despite using half the training size that is used for the BlenderBot 90M and Seq2Seq models.

Performance vs. amount of data or context size

Taking DialoGPT, the best model from Table 5.5, and doing ablation studies over both the training set size and the context size, we arrive at results in Tables 5.6 and 5.7, respectively. Increasing the training set size by doubling the number of dialogue turns brings improvement by lowering the perplexity for the model of each language. However, doubling the context size, does not result in a similar effect. Perplexity only improves

[13] slack.com

Table 5.5: Results for the 3 main models (c14: context size 14; sd: standard deviation; Hausa seq2seq appears to overfit) (Adewumi et al., 2022a).

Language	**Model**	**Training turns**	**Perplexity**	
			Dev (sd)	Test (sd)
Pidgin English	DialoGPT c14	500	67.57 (2.53)	90.18 (3.24)
	BlenderBot 90M	1,000	81.23 (0)	81.23 (0)
	Seq2Seq	1,000	277.2 (15)	277.2 (15)
Yorùbá	DialoGPT c14	500	12.63 (0.47)	10.66 (0.40)
	BlenderBot 90M	1,000	154.43 (0.06)	154.43 (0.06)
	Seq2Seq	1,000	45.85 (1.41)	45.85 (1.41)
Hausa	DialoGPT c14	500	26.40 (0.75)	35.95 (0.73)
	BlenderBot 90M	1,000	39.39 (1.61)	39.39 (1.61)
	Seq2Seq	1,000	1.92 (0.12)	1.92 (0.12)
Wolof	DialoGPT c14	500	15.2 (0.09)	26.41 (0.10)
	BlenderBot 90M	1,000	108.7 (0)	108.7 (0)
	Seq2Seq	1,000	401.6 (10.39)	401.6 (10.39)
Swahili	DialoGPT c14	500	20.03 (0.29)	17.02 (0.22)
	BlenderBot 90M	1,000	128.8 (0.10)	128.8 (0.10)
	Seq2Seq	1,000	134.5 (2.75)	134.5 (2.75)
Kinyarwanda	DialoGPT c14	500	24.47 (0.17)	26.45 (0.17)
	BlenderBot 90M	1,000	177.87 (0.06)	177.87 (0.06)
	Seq2Seq	1,000	195.07 (7.66)	195.07 (7.66)

when we half the context size from 14 to 7. The results are statistically significant. P-values ($p < 0.0001$) for the difference of two means of the two-sample t-test (between the two lowest results) for all the languages are smaller than alpha (0.05). Given that these results are obtained with small data, increasing the data size will improve the results.

Human evaluation

Table 5.8 shows that the single-turn dialogues of the Nigerian Pidgin English are human-like 78.1% of the time by majority votes. 34.4% of them are unanimously judged as human-like, which is higher than both the 3-way tie (when each annotator voted for each different category) of 15.6% or non-human-like of 6.3%. This is very likely because Nigerian Pidgin English is closely related to the English language, which is the language of pretraining. Meanwhile, the Yorùbá transcript has 0% human-like single-turn conversation. This may be because of the language's morphology and written accent, among others reasons. It has the most peculiarities in written form, as shown in Table **??**, making it challenging for the model. Wolof, Hausa, Swahili and Kinyarwanda follow after Nigerian Pidgin English with 65.6%, 31.3%, 28.1% and 28.1% of conversations assessed as human-like, respectively.

The Fleiss Kappa (k) scores are not interpretable using the Kappa 2 annotators on 2 classes guide (Landis and Koch, 1977), since this study uses 3 annotators on 3 classes

Table 5.6: Ablation study of DialoGPT-c7 over training turns (c7: context size 7; sd: standard deviation; bold figures are the better values per language) (Adewumi et al., 2022a)

Language	Training turns	Perplexity	
		Dev (sd)	Test (sd)
Nigerian Pidgin English	500	42.55 (0)	52.81 (0)
	1,000	**37.95** (0.66)	**46.56** (1.13)
Yorùbá	500	10.52 (0.04)	9.65 (0.01)
	1,000	**7.22** (0.06)	**8.76** (0.08)
Hausa	500	18.53 (0.23)	25.7 (0.4)
	1,000	**9.92** (0.05)	**12.89** (0.04)
Wolof	500	15.2 (0.09)	26.41 (0.10)
	1,000	**14.91** (0.3)	**25.85** (0.04)
Swahili	500	15.55 (0.17)	14.22 (0.14)
	1,000	**9.63** (0)	**9.36** (0.03)
Kinyarwanda	500	19.28 (0.19)	21.62 (0.22)
	1,000	**10.85** (0)	**14.18** (0.08)

Table 5.7: Ablation study of DialoGPT over context sizes for training set with 1,000 turns (c7, c14: context sizes 7 & 14; sd: standard deviation; bold figures are the better values per language) (Adewumi et al., 2022a)

Language	Context size	Perplexity	
		Dev (sd)	Test (sd)
Nigerian Pidgin English	c7	**37.95** (0.66)	**46.56** (1.13)
	c14	70.21 (2.17)	92.23 (2.33)
Yorùbá	c7	**7.22** (0.06)	**8.76** (0.08)
	c14	7.63 (0.13)	9.11 (0.14)
Hausa	c7	**9.92** (0.05)	**12.89** (0.04)
	c14	11.30 (0.04)	15.16 (0.05)
Wolof	c7	**14.91** (0.3)	**25.85** (0.04)
	c14	16.61 (0.2)	30.37 (0.08)
Swahili	c7	**9.63** (0)	**9.36** (0.03)
	c14	11.07 (0.04)	10.71 (0.05)
Kinyarwanda	c7	**10.85** (0)	**14.18** (0.08)
	c14	12.84 (0.1)	17.43 (0.14)

and k is lower when the classes are more (Sim and Wright, 2005). This study confirms the observation made by Gwet (2014) that the interpretation guide may be more harmful than helpful. Perez Almendros et al. (2020) report how the Kappa score rose from 41% to 61% when the classes were reduced from three to two. CUS is apparently a more intuitive and reliable method, at least, in this case because of the homogeneous test conversations. Figure 5.9 shows the human-likeness scores and the CUS in a bar chart for the languages, as given in Table 5.8.

Table 5.8: Human evaluation results of 3 annotators on 3 classes using single-turn conversations. A recent human-human upperbound is 92.1%, according to Adewumi et al. (2022c). The subjective Kappa example of 2 annotators on 2 classes does not apply here since Kappa is lower when classes are more (Sim and Wright, 2005). - implies not applicable (Adewumi et al., 2022a).

Model language	**Scale (majority votes - 2/3)**				**CUS**	**Fliess *k***
	H (%)	U (%)	N (%)	3-way (%)	%	
Nigerian Pidgin English	78.1	0	6.3	15.6	66.7	-0.079
Yorùbá	0	3.1	75	21.9	33.3	-0.154
Hausa	31.3	6.3	53.1	9.4	66.7	0.228
Wolof	65.6	0	31.3	3.1	100	0.070
Swahili	28.1	15.6	34.4	21.9	66.7	0.067
Kinyarwanda	28.1	25	34.4	12.5	66.7	0.091
	unanimous votes - 3/3					
Nigerian Pidgin English	34.4	0	0	-	66.7	
Yorùbá	0	0	25	-	33.3	
Hausa	12.5	0	21.9	-	66.7	
Wolof	15.6	0	9.4	-	100	
Swahili	9.4	0	9.4	-	66.7	
Kinyarwanda	9.4	0	6.3	-	66.7	

Error analysis

A few of the challenges noticeable with the models are highlighted here. DialoGPT, though the best model in the investigations, shows particular weakness with the Yorùbá language by failing to generate the right word morphology, sometimes. Below are 2 examples where the conversations are not coherent.

- *prompt:* Ẹsẹ̀ ń ro mí
- *response:* Eèlóò ni orúkọ ọdún?
- *prompt:* Ìyẹn náà da, wọ́n ṣè ń darúgbó náà ni
- *response:* Kì ni orúkọ́ orúnju rẸ?

BlenderBot 90M mixes English responses with its generation atimes, even though the prompts are of the foreign target language. Two examples of prompts in Hausa and the generated responses are given below. One can observe that it echoes the prompt given to it in the first pair of example.

- *prompt:* Barka dai, me za ku iya gaya mani game da gidan abincin bangkok na burni?
- *response:* barka dai , za za ku iya gaya mani game da gidan abincin bangkok na burni ? i ' m not sure if you ' re being sarcastic or not .

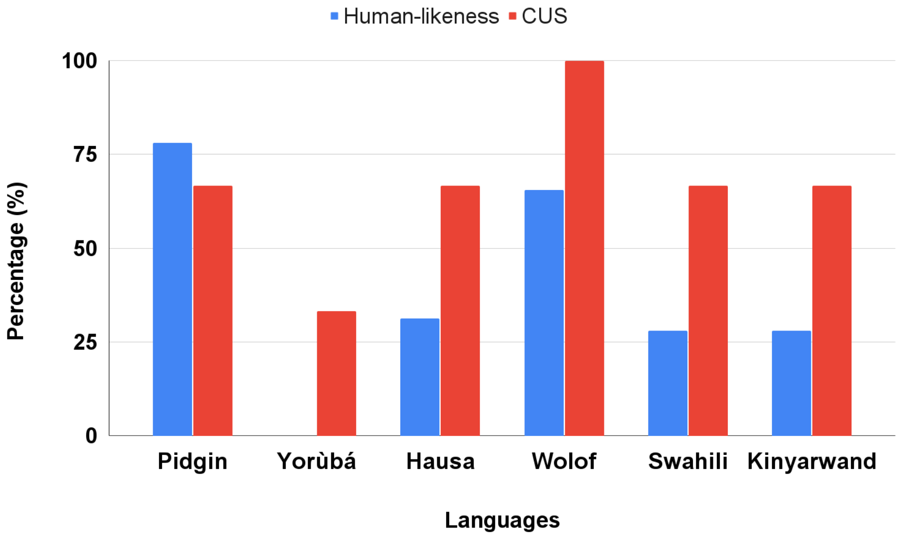

Figure 5.9: Bar chart of human-likeness scores and CUS for the languages (Adewumi et al., 2022a).

- *prompt:* Ee na mutum 2 da 19:30 na ranar Lahadi don Allah. Ina bukatan lambar nuni kuma.
- *response:* haha , i ' m not sure if i ' m the only one who thinks that .

A big drawback of the seq2seq model is the repetitive nature of some of its responses across different prompts. Two examples of the prompts and generated responses for the Nigerian Pidgin English are given below. Exactly the same response is generated for two different prompts and the response is not coherent.

- *prompt:* I need one place to chop for the center wey dey expensive. Good bye
- *response:* I dey go like .
- *prompt:* e dey okay. i fit get that number. you fit also recommend me hotel wey dey expensive
- *response:* I dey go like .

Evaluator feedback

In post-evaluation feedback, some evaluators explained that coherence of the conversation and the grammar mattered to them in making their judgment. For example, considering

Yorùbá, responses that referenced inanimate objects as if in the context of animate objects or humans were voted as non-human-like. For Wolof, many conversations are human-like but for cases where the responses were inconsistent with the prompt given. For example, there were conversations that were hard for an evaluator to judge because the responses are questions to question-prompts. Such conversations were awarded the uncertain (U) votes by the evaluator.

Chapter 6

Conclusion and Future Work

"Good night toiling, good sleep."

(Apostrophe)

In the end, users tend to want intelligent systems with the ability to communicate in natural language. The more original and colourful the communication, the better. The experiences the pioneering chatbot ELIZA, by Weizenbaum (1969), created with her users suggest this might be the case. The road to it might seem long but it may be possible in the foreseeable future. One important element to achieving human-like conversations will be to endow conversational models with idiom-awareness since a conversational system that can respond in a similar way to its user, in figurative speech, is more fitting, as this study shows.

6.1 Conclusion

This thesis confirms two important hypotheses about open-domain conversational systems that are idiom-aware and deep monolingual models. For the confirmation of the first hypothesis that an open-domain conversational system that is idiom-aware, generates more fitting responses to prompts containing idioms, Chapter 2 introduced the PIE-English idioms corpus. Chapter 4 presented results of training the SoTA DialoGPT model on the corpus. The PIE-English idioms corpus offers opportunities for further research, as the dataset may be adapted or expanded in different ways. It may not be sufficient to train models on data that exclude idioms and it may not always be practical to substitute idioms with their literal meaning in exchanges between users and the conversational systems. Instead, careful curation of figurative language data is essential to train open-domain conversational deep learning models. This is because idioms or figurative language is part and parcel of many human languages and cannot be ignored if we must achieve the rich conversation that is typical of natural languages with conversational systems.

For the confirmation of the second hypothesis that deep monolingual models learn some abstractions that generalise across languages, Chapter 5 presented results of transferability from English to seven other diverse languages. Some of the abstractions seem to be the linguistic universals, which are common across many languages. They are semantic similarity across languages through polysemous words (Youn et al., 2016) and minimal dependency length (MDL) (Futrell et al., 2015). Out of the seven languages, for which this hypothesis is demonstrated, the only one (Yorùbá) that seems not to fit the hypothesis, based on human evaluation, may actually do so if better quality data, such as the MultiWOZ, is used. The linguistic universals in languages reveal that though we humans are so diverse, we are also very similar in many ways.

Four important research questions (RQ) are addressed in this thesis: 1) How importantly do hyper-parameters influence word embeddings' performance? 2) What factors are important for developing ethical and robust conversational systems? 3) To what extent can models trained on figures of speech (idioms) enhance NLP? And 4) How can models trained on figures of speech (idioms) enhance open-domain, data-driven chatbots for robust assistance? The following contributions arose as the outcome of addressing the hypotheses and RQs.

1. The Swedish analogy test set for evaluating Swedish word embeddings is created and released publicly under the CC-BY4 licence. The resource, which was verified by Språkbanken, is hosted on the Språkbanken website[1].

2. The Potential Idiomatic Expression (PIE)-English idioms corpus, is created and released publicly under the CC-BY4 licence. The purpose of the corpus is to train ML models in idiom identification and classification. This resource is hosted on the International Conference on Language Resources and Evaluation (LREC) platform[2].

3. The AfriWOZ dialogue dataset of parallel corpora of 6 African languages is created and released under the CC-BY4 licence. This dataset is primarily for training open-domain conversational systems but it may easily be adapted for other relevant NLP tasks, like MT, automatic speech recognition (ASR), and task-based conversational systems. The resource is hosted online[3].

4. Credibility unanimous score (CUS) is introduced for measuring IAA of conversation transcripts. The assumption behind CUS is simple and provides advantages over some other methods, such as Fleiss Kappa (k), because it seems more intuitive, easier to calculate (as it is based on percentages), and seemingly less sensitive to changes in the number of categories being evaluated. Besides, the homogeneous samples are also used to test the credibility of the annotators and determine majority agreement on the human-human (or homogeneous) conversations in the transcript.

[1]spraakbanken.gu.se/en/resources/analogy
[2]lrec2022.lrec-conf.org/en/
[3]github.com/masakhane-io/chatbots-african-languages

5. We show insights into energy-saving and time-saving benefits of more optimal embeddings from better hyperparameter combinations and relatively smaller corpora.

6. Selected word embeddings in English, Swedish and Yorùbá are created and released for public access.

7. The codes used in this work are made open-source and hosted on Github[4], under the CC-BY4 licence.

8. The model checkpoints developed in the course of this thesis are made available on the HuggingFace hub[5].

9. The philosophical argument for developing robust and ethical conversational systems are raised and may serve as a springboard for further helpful discussions around the subject.

Furthermore, the importance of ethics in the development of open-domain conversational systems cannot be over-emphasised. Privacy concerns, offensive/hateful messages, and harmful bias of all kinds are some of the issues that should be considered (Jurafsky and Martin, 2020). The use of model cards and data statements are some of the ways to address these concerns, though they should not be taken as exoneration from responsibility. This thesis provides such model cards and data statements for the deep models in this work, especially since the pretraining data are from online public sources that are known to contain all kinds of views (including undesirable ones) and suffer from the concerns already identified.

6.2 Future work

This work has provided some resources and insight into open-domain conversational systems but there are still existing challenges and many possibilities to be explored. The Swedish analogy test set could be extended and made balanced across all the subsections. This may provide a more robust evaluation of Swedish embeddings though intrinsic evaluations are known to have shortcomings (Chiu et al., 2016). The PIE-English idioms corpus may be adapted or extended by increasing the samples for the classes with very little samples or increasing the number of classes that are represented. Doing so may produce more fitting responses from open-domain conversational systems. In addition, investigating and designing better decoding algorithms that will be much similar to the distribution of human conversation will make achieving human-like conversations realistic (Holtzman et al., 2020).

Since this may be the first thesis exploring cross-lingual transferability from deep monolingual English models to low-resource languages for open-domain conversational systems, scaling up this work to more languages will establish the extent to which

[4]github.com/tosingithub
[5]huggingface.co/tosin

the hypothesis holds. Transfer learning, based on pretrained deep models, provides energy-saving and time-saving benefits for downstream tasks when finetuning is applied. Zero/Few-shot learning provides gains in this regard also and may be advantageous for low-resource languages. AfriWOZ may provide the opportunity to develop open-domain conversational systems that can chat with each other (in machine-machine conversations), thereby continually generating high-quality data for low-resource languages. The automatically generated data may be useful for other NLP tasks such as automatic speech recognition (ASR), NER, MT, task-based conversational AI, and automatic text summarisation, among others. The future holds many possibilities and it's crucial to continue to have discussions, whether philosophical or practical, in order to shape the future for ethical and robust open-domain conversational systems.

Appendices

A Appendix A

Data statement for the Swedish analogy test set for evaluating Swedish word embeddings.

	Details
Curation rationale	Due to the unavailability of Swedish evaluation dataset for word embeddings this analogy test set was created.
Dataset language	Swedish
	Demographics of contributors
No of contributors	1
Age	42
Gender	Male
Language	L2
	Demographics of annotators
No of annotators	2
	Annotator 1
Age	-
Gender	Male
Language	L1
	Annotator 2
Age	-
Gender	Male
Language	L1
	Data characteristics
Total samples	20,637
Number of Sections	2 Main sections
Semantic section	10,380 samples (5 sections- capital-common-countries (342), capital-world (7.832), currency (42), city-in-state (1,892), family (272))
Syntactic section	10,257 samples (6 sections - gram2-opposite (2,652), gram3-comparative (2,162), gram4-superlative (1,980), gram6-nationality-adjective (12), gram7-past-tense (1,891), gram8-plural (1,560))
	Others
IAA	98.93% (raw percentage)
Licence	CC-BY 4.0.

Table 6.1:

B Appendix B

Data statement for the PIE-English idioms corpus for idiom identification.

	Details
Curation rationale	Due to the unavailability of idioms dataset with more than the 2classes of literal & general figurative speech classification, this dataset was created.
Dataset language	English
	Demographics of contributors
No of contributors	4
Age	42 | - | - | -
Gender	Male | Female | Female | Female
Language	L2 | L2 | L2 | L2
	Demographics of annotators
No of annotators	2
	Annotator 1
Age	-
Gender	Male
Language	L2
	Annotator 2
Age	-
Gender	-
Language	L2
	Data characteristics
Total samples	20,174
Number of classes	10
Number of cases	1,197 (e.g. *"the nick of time"*, *"a laugh a minute"*)
	Total samples of euphemism (2,384), literal (1,140), metaphor (14,666), personification (448), simile (1,232), parallelism (64), paradox (112), hyperbole (48), oxymoron (48), and irony (32)
Base data	BNC and UKWaC.
	Others
IAA	88.89% (raw percentage)
Licence	CC-BY 4.0.

Table 6.2:

C Appendix C

Data statement for the AfriWOZ dataset for open-domain dialogue & other NLP models.

	Details
Curation rationale	Due to the unavailability of dialogue data for low-resource African languages, this dataset was created.
Dataset language	Swahili, Wolof, Hausa, Nigerian Pidgin English, Kinyarwanda & Yorùbá
	Demographics of contributors
No of contributors	19
Age	-
Gender	Male & Female
Language	L1
	Demographics of annotators
No of annotators	Not applicable
	Data characteristics
Total samples	1,500 turns per language
Total natural languages	6 (Swahili, Wolof, Hausa, Nigerian Pidgin English, Kinyarwanda & Yorùbá)
Training set turns per language	1,000
Validation set turns per language	250
Test set turns per language	250
Domains covered	hotel, restaurant, taxi and booking.
Base data	MultiWOZ and 2 blogs for Yorùbá only.
	Others
IAA	CUS 33.3% - 100%
Licence	CC-BY 4.0.

Table 6.3:

D Appendix D

Model card for DialoGPT, trained on MultiWOZ

	Details
Model rationale	This is a fine-tuned model of DialoGPT (medium) on the MultiWOZ v2.2 dataset. It is intended to be used as a conversational system.
Limitation	The dataset it's trained on is limited in scope, as it covers only certain domains such as restaurants, hotel, taxi, train, hospital and police.
Result	The perplexity achieved on the test and validation sets are 10.25 and 10.71, respectively.
	Others
Model link	https://huggingface.co/tosin/dialogpt_mwoz
More	Information about the original pre-trained model can be found at https://huggingface.co/microsoft/DialoGPT-medium
Licence	CC-BY 4.0.
	Generation example from an interactive environment:
Role	**Response**
User	I would like to have breakfast.
Bot	There are several options for Chinese in the city centre. Are you looking for a particular price range?
User	I'm looking for free food
Bot	There are 4 Chinese restaurants that are in the cheap price range. 3 in centre and one in north.
User	Not good enough.
Bot	How about Charlie Chan? It is a Chinese restaurant in the centre of town.

Table 6.4:

E Appendix E

Model card for DialoGPT, trained on Gothenburg Dialogue Corpus (GDC)

	Details
Model rationale	This is a fine-tuned model of the DialoGPT (medium) on the Swedish Gothenburg Dialogue Corpus (GDC). It is intended to be used as a Swedish conversational system.
Limitation	The GDC dataset it's trained on is limited in scope, as it's from the transcription of dialogues of about 25 different social activities, including debates, academic seminars and situations. The dataset has about 108K turns.
Result	The perplexity achieved on the test and validation sets (from 80:10:10 split) are 27.81 and 27.81, respectively.
	Others
Model link	https://huggingface.co/tosin/dialogpt_sv
More	Information about the original pre-trained model can be found at https://huggingface.co/microsoft/DialoGPT-medium
Licence	CC-BY 4.0.
	Generation example from an interactive environment:
Role	**Response**
User	Jag ska fika.
Bot	ja det är ju så att jag kanske inte kan få den att göra det
User	tack så mycket
Bot	tack tack ses vi om sex vecko igen

Table 6.5:

F Appendix F

Model card for DialoGPT, trained on the PIE-English idioms corpus

	Details
Model rationale	This is a fine-tuned model of DialoGPT (medium)-MultiWOZ v2.2 on the Potential Idiomatic Expression (PIE)-English idioms corpus. It is intended to be used as an idiom-aware conversational system.
Limitation	The dataset it's trained on is limited in scope, as it covers only 10 classes including metaphor, simile, euphemism, parallelism, personification, oxymoron, paradox, hyperbole, irony and literal.
Result	The perplexity achieved on the test and validation sets are 160.9 and 160.46, respectively.
	Others
Model link	https://huggingface.co/tosin/dialogpt_mwoz_idioms
More	Information about the original pre-trained model can be found at https://huggingface.co/tosin/dialogpt_mwoz
Licence	CC-BY 4.0.
	Generation example from an interactive environment:
Role	**Response**
User	remind her to take it off the hook.
Bot	the bookie took it on the nod
User	A lengthy break is in order ; so make sure you keep your nose clean
Bot	they are under a cloud because of fears about US President's wife Hillary Clinton
User	Some families breed like rabbits and send us all their male heirs
Bot	he was out for the count

Table 6.6:

G Appendix G

Model card for DialoGPT, trained on the wolof data in the AfriWOZ dataset

	Details
Model rationale	This is a fine-tuned model of DialoGPT (small) on the Wolof data of the AfriWOZ dataset. It is intended to be used as a conversational system in Wolof language.
Limitation	The dataset it's trained on is limited in scope, as it covers only certain domains such as restaurants, hotel, taxi, and booking.
Result	The perplexity achieved on the test and validation sets are 20.74 and 11.66, respectively.
	Others
Model link	https://huggingface.co/tosin/dialogpt_afriwoz_wolof
More	Information about the original pre-trained model can be found at https://huggingface.co/microsoft/DialoGPT-small
Licence	CC-BY 4.0.
	Generation example from an interactive environment:
Role	**Response**
User	sowwu bi la bëg
Bot	waw, ndax mën ngan ma jox ay leeral si bërëb buñu feetee ak lekkukaay bi
User	soxla na dem gawu. ngir leeral rek bëg na joggee peterborough di dem cambridge
Bot	soxlu na dem ginaw jurrom ñenti waxtu ak ñaar fukki simili ak jurrom ben.
User	waw jiss sal ma ben bu seer si njëg si toluway
Bot	jënd bi baaxna. jerejef. ba benen yoon.

Table 6.7:

References

D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9: 1116–1131, 2021.

T. Adewumi, M. Adeyemi, A. Anuoluwapo, B. Peters, H. Buzaaba, O. Samuel, A. M. Rufai, B. Ajibade, T. Gwadabe, M. M. K. Traore, T. Ajayi, S. Muhammad, A. Baruwa, P. Owoicho, T. Ogunremi, P. Ngigi, O. Ahia, R. Nasir, F. Liwicki, and M. Liwicki. Ìtàkúròso: Exploiting cross-lingual transferability for natural language generation of dialogues in low-resource, african languages. 2022a. doi: 10.48550/ARXIV.2204.08083. URL https://arxiv.org/abs/2204.08083.

T. Adewumi, L. Alkhaled, H. Alkhaled, F. Liwicki, and M. Liwicki. Ml_ltu at semeval-2022 task 4: T5 towards identifying patronizing and condescending language. *arXiv preprint arXiv:2204.07432*, 2022b.

T. Adewumi, R. Brännvall, N. Abid, M. Pahlavan, S. S. Sabry, F. Liwicki, and M. Liwicki. Småprat: Dialogpt for natural language generation of swedish dialogue by transfer learning. In *5th Northern Lights Deep Learning Workshop, Tromsø, Norway*, volume 3. Septentrio Academic Publishing, 2022c. doi: https://doi.org/10.7557/18.6231.

T. Adewumi, F. Liwicki, and M. Liwicki. Word2vec: Optimal hyperparameters and their impact on natural language processing downstream tasks. *Open Computer Science*, 12 (1):134–141, 2022d. doi: doi:10.1515/comp-2022-0236. URL https://doi.org/10.1515/comp-2022-0236.

T. P. Adewumi. Inner loop program construct: A faster way for program execution. *Open Computer Science*, 8(1):115–122, 2018. doi: doi:10.1515/comp-2018-0004. URL https://doi.org/10.1515/comp-2018-0004.

T. P. Adewumi, F. Liwicki, and M. Liwicki. Conversational systems in machine learning from the point of view of the philosophy of science—using alime chat and related studies. *Philosophies*, 4(3):41, 2019.

T. P. Adewumi, F. Liwicki, and M. Liwicki. The challenge of diacritics in yoruba embeddings. *arXiv preprint arXiv:2011.07605*, 2020a.

T. P. Adewumi, F. Liwicki, and M. Liwicki. Corpora compared: The case of the swedish gigaword & wikipedia corpora. *arXiv preprint arXiv:2011.03281*, 2020b.

T. P. Adewumi, F. Liwicki, and M. Liwicki. Exploring swedish & english fasttext embeddings for ner with the transformer. *arXiv preprint arXiv:2007.16007*, 2020c.

T. P. Adewumi, R. Vadoodi, A. Tripathy, K. Nikolaidou, F. Liwicki, and M. Liwicki. Potential idiomatic expression (pie)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*, 2021.

D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020. doi: 10.48550/arXiv.2001.09977.

C. C. Aggarwal and C. Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

J. Alabi, K. Amponsah-Kaakyire, D. Adelani, and C. España-Bonet. Massive vs. curated embeddings for low-resourced languages: the case of yorùbá and twi. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2754–2762, 2020.

L. Alexander and M. Moore. Deontological ethics. 2007.

J. Allwood, L. Grönqvist, E. Ahlsén, and M. Gunnarsson. Annotations and tools for an activity based spoken language corpus. In *Current and new directions in discourse and dialogue*, pages 1–18. Springer, 2003. doi: 10.1007/978-94-010-0019-2_1.

C. Alm-Arvius. *Figures of speech*. Studentlitteratur, 2003.

M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421.

K. Bach and R. M. Harnish. Linguistic communication and speech acts. 1979.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations, ICLR 2015*, 2015. doi: 10.48550/arXiv.1409.0473. URL https://arxiv.org/pdf/1409.0473.pdf.

E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003.

Y. Bengio, Y. LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

J. Birke and A. Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Y. Bizzoni, S. Chatzikyriakidis, and M. Ghanimifard. "deep" learning : Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, Copenhagen, Denmark, Sept. 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-4906. URL https://aclanthology.org/W17-4906.

Y. Bizzoni, S. Chatzikyriakidis, and M. Ghanimifard. "deep" learning: Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, 2017b.

B. Björkman. Language ideology or language practice? an analysis of language policy documents at swedish universities. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, 33(3-4):335–363, 2014.

D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd. Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173, 1977.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva. The groningen meaning bank. In *Handbook of linguistic annotation*, pages 463–496. Springer, 2017.

L. Bradeško and D. Mladenić. A survey of chatbot systems through a loebner prize competition. In *Proceedings of Slovenian language technologies society eighth conference of language technologies*, pages 34–37. Institut Jožef Stefan Ljubljana, Slovenia, 2012.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

M. Brysbaert and M. Stevens. Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition*, 1(1), 2018.

A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*, volume 2, pages 2–2, 2001.

P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547.

B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, B. Goodrich, D. Duckworth, S. Yavuz, A. Dubey, K.-Y. Kim, and A. Cedilnik. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1459. URL https://aclanthology.org/D19-1459.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

B. Chiu, A. Korhonen, and S. Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, pages 1–6, 2016.

N. Chomsky. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124, 1956.

K. Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

A. Clark, C. Fox, and S. Lappin. *The handbook of computational linguistics and natural language processing*, volume 118. John Wiley & Sons, 2012.

K. M. Colby, S. Weber, and F. D. Hilf. Artificial paranoia. *Artificial Intelligence*, 2(1): 1–25, 1971.

K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3:199–221, 1972.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

P. Cook, A. Fazly, and S. Stevenson. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48, 2007.

A. P. Cowie and R. Mackin. Oxford dictionary of current idiomatic english v. 2:phrase, clause & sentence idioms. 1983.

M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2 (1):1–24, 2015.

R. Creath. Logical empiricism. 2011.

D. Crystal. *The Cambridge encyclopedia of the English language*. Cambridge university press, 2018.

B. V. Dasarathy. Nearest neighbor (nn) norms: Nn pattern classification techniques. *IEEE Computer Society Tutorial*, 1991.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018a.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Multilingual bert. 2018b.

M. Diab and P. Bhutada. Verb noun construction mwe token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, pages 17–22, 2009.

E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.656. URL https://aclanthology.org/2020.emnlp-main.656.

P. Drew and E. Holt. Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in society*, pages 495–522, 1998.

J.-B. Du Prel, G. Hommel, B. Röhrig, and M. Blettner. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 106(19):335, 2009.

J. Elkner, A. B. Downey, and C. Meyers. How to think like a computer scientist: Learning with python documentation. *Release*, 2010.

D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Artificial Intelligence and Statistics*, pages 153–160. PMLR, 2009.

D. Erhan, A. Courville, Y. Bengio, and P. Vincent. Why does unsupervised pre-training help deep learning? In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 201–208, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/erhan10a.html.

M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, and D. Hakkani-Tür. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. 2019.

M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, A. Kumar, A. Goyal, P. Ku, and D. Hakkani-Tur. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.53.

K. Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012.

P. Fallgren, J. Segeblad, and M. Kuhlmann. Towards a standard dataset of swedish word vectors. In *Sixth Swedish Language Technology Conference (SLTC), Umeå 17-18 nov 2016*, 2016.

M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://aclanthology.org/E14-1049.

M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2506. URL https://aclanthology.org/W16-2506.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131, 2002.

J. R. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

W. T. Fitch. Unity and diversity in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1563):376–388, 2011.

L. Fleck. *Genesis and development of a scientific fact*. University of Chicago Press, 2012.

J. Forlizzi, J. Zimmerman, V. Mancuso, and S. Kwak. How interface agents affect interaction between humans and computers. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces*, pages 209–221, 2007.

E. S. Foundation and A. E. Academies. *The European code of conduct for research integrity*. European Science Foundation, 2017.

R. M. French and C. Labiouse. Four problems with extracting human semantics from large text corpora. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24, 2002.

R. W. Frick. Accepting the null hypothesis. *Memory & Cognition*, 23(1):132–138, 1995.

R. W. Frick. The appropriate use of null hypothesis testing. *Psychological Methods*, 1 (4):379, 1996.

A. D. Friederici. *Language in our brain: The origins of a uniquely human capacity*. MIT Press, 2017.

J. Fuegi and J. Francis. Lovelace & babbage and the creation of the 1843'notes'. *IEEE Annals of the History of Computing*, 25(4):16–26, 2003.

R. Futrell, K. Mahowald, and E. Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112 (33):10336–10341, 2015. doi: 10.1073/pnas.1502134112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1502134112.

P. Gage. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

G. Galilei. Discourses and mathematical demonstrations relating to two new sciences. *Leiden (1638)*, 1954.

V. Gangal, H. Jhamtani, E. Hovy, and T. Berg-Kirkpatrick. Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4079–4090, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.357. URL https://aclanthology.org/2021.findings-acl.357.

A. Gatt and E. Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61: 65–170, 2018.

S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Aremu, A. Bosselut, K. R. Chandu, M.-A. Clinciu, D. Das, K. Dhole, W. Du, E. Durmus, O. Dušek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, M. Kale, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, A. Niyongabo Rubungo, S. Osei, A. Parikh, L. Perez-Beltrachini, N. R. Rao, V. Raunak, J. D. Rodriguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. Sobrevilla Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

K. Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

L. Grant and L. Bauer. Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied linguistics*, 25(1):38–61, 2004.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*, 2018.

C. Grover and R. Tobin. A gazetteer and georeferencing for historical english documents. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 119–127, 2014.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2), 2012.

K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC, 2014.

H. Haagsma, J. Bos, and M. Nissim. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 279–287, 2020.

G. Hackeling. *Mastering Machine Learning with scikit-learn.* Packt Publishing Ltd, 2017.

M. T. Hagan, H. B. Demuth, and M. Beale. *Neural network design.* PWS Publishing Co., 1997.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

T. B. Hashimoto, H. Zhang, and P. Liang. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1169. URL https://aclanthology.org/N19-1169.

K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

B. Heine, D. Nurse, et al. *African languages: An introduction*. Cambridge University Press, 2000.

P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2018.

D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2712–2721. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hendrycks19a.html.

S. Hengchen and N. Tahmasebi. SuperSim: a test set for word similarity and relatedness in Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 268–275, Reykjavik, Iceland (Online), May 31–2 June 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.27.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations, ICLR 2020*, 2020. URL https://arxiv.org/pdf/1904.09751.pdf.

E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191, 2020.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for NLP. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Z. Hu, H. Shi, B. Tan, W. Wang, Z. Yang, T. Zhao, J. He, L. Qin, D. Wang, X. Ma, et al. Texar: A modularized, versatile, and extensible toolkit for text generation. *arXiv preprint arXiv:1809.00794*, 2018.

K. U. Ihemere. A basic description and analytic treatment of noun clauses in nigerian pidgin. *Nordic journal of African studies*, 15(3):296–313, 2006.

N. Indurkhya and F. J. Damerau. *Handbook of natural language processing*. Chapman and Hall/CRC, 2010.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, sep 1999. ISSN 0360-0300. doi: 10.1145/331499.331504. URL https://doi.org/10.1145/331499.331504.

S. Javed, T. P. Adewumi, F. S. Liwicki, and M. Liwicki. Understanding the role of objectivity in machine learning and research evaluation. *Philosophies*, 6(1):22, 2021.

G. Jefferson. Side sequences. *Studies in social interaction*, 1972.

H. Jhamtani, V. Gangal, E. Hovy, and T. Berg-Kirkpatrick. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.592. URL https://aclanthology.org/2021.emnlp-main.592.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Limited, 2020. ISBN 9789332518414. URL https://books.google.se/books?id=ZalcjwEACAAJ.

L. J. Kazmier. *Theory and problems of business statistics*. McGraw-Hill, 2004.

A. Kerry, R. Ellis, and S. Bull. Conversational agents in e-learning. In *International conference on innovative techniques and applications of artificial intelligence*, pages 169–182. Springer, 2008.

S. C. Kleene et al. Representation of events in nerve nets and finite automata. *Automata studies*, 34:3–41, 1956.

D. Klein and C. D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pages 478–485, 2004.

E. Konig and J. Van der Auwera. *The germanic languages*. Routledge, 2013.

I. Korkontzelos, T. Zesch, F. M. Zanzotto, and C. Biemann. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, 2013.

K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

T. S. Kuhn. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press, 1970.

S. G. Kwak and J. H. Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144, 2017.

G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago press, 2008.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

M. Lee, S. Ackermans, N. Van As, H. Chang, E. Lucas, and W. IJsselsteijn. Caring for vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225, 2015.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

L. Li and C. Sporleder. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323, 2009.

Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1099.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

Q. Liu, M. J. Kusner, and P. Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020a.

Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020b.

H. E. Longino. *Science as social knowledge*. Princeton university press, 2020.

M. M. Louwerse, A. C. Graesser, S. Lu, and H. H. Mitchell. Social cues in animated conversational agents. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(6):693–704, 2005.

A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256, 2015.

M. Lundell Vinkler and P. Yu. Conversational chatbots with memory-based question and answer generation, 2020.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

A. Maedche. Gender bias in chatbot design. *Chatbot Research and Design*, page 79, 2020.

C. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

R. Mao, C. Lin, and F. Guerin. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1113. URL https://aclanthology.org/P18-1113.

G. Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

M. L. Mauldin. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, volume 94, pages 16–21, 1994.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

A. S. Medearis. *Too Much Talk: A West African Folktale.* Candlewick, 1995.

W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1008.

A. Miller, W. Feng, D. Batra, A. Bordes, A. Fisch, J. Lu, D. Parikh, and J. Weston. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2014. URL https://aclanthology.org/D17-2014.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

T. M. Mitchell et al. Machine learning, 1997.

S. Mohammad, E. Shutova, and P. Turney. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-2003. URL https://aclanthology.org/S16-2003.

F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In *International workshop on artificial intelligence and statistics*, pages 246–252. PMLR, 2005.

B. M. Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

C. I. Nass and S. Brave. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press Cambridge, 2005.

G. Neff and P. Nagy. Automation, algorithms, and politics| talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:17, 2016.

W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. Muhammad, S. Kabongo Kabenamualu, S. Osei, F. Sackey, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. M. Berhe, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, M. Espoir, E. van Biljon, D. Whitenack, C. Onyefuluchi, C. C. Emezue, B. F. P. Dossou, B. Sibanda, B. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin, and A. Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL https://aclanthology.org/2020.findings-emnlp.195.

I. Newton. *Philosophiae naturalis principia mathematica*, volume 1. G. Brookman, 1833.

R. S. Nickerson. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2):241, 2000.

O. Olabiyi and E. T. Mueller. Multiturn dialogue response generation with autoregressive transformer models. *arXiv preprint arXiv:1908.01841*, 2019.

X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://www.aclweb.org/anthology/P17-1178.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

M. Paquette, E. J. Sommerfeldt, and M. L. Kent. Do the ends justify the means? dialogue, development communication, and deontological ethics. *Public Relations Review*, 41(1):30–39, 2015.

J. Peng, A. Feldman, and H. Jazmati. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 507–511, Hissar, Bulgaria, Sept. 2015a. INCOMA Ltd. Shoumen, BULGARIA. URL https://aclanthology.org/R15-1066.

J. Peng, A. Feldman, and H. Jazmati. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 507–511, 2015b.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014a. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014b.

C. Perez Almendros, L. Espinosa Anke, and S. Schockaert. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.518. URL https://aclanthology.org/2020.coling-main.518.

C. Pérez-Almendros, L. Espinosa-Anke, and S. Schockaert. SemEval-2022 Task 4: Patronizing and Condescending Language Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, Oct.-Nov. 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1179. URL https://aclanthology.org/D18-1179.

J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.7.

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

E. C. Polomé. Swahili language handbook. 1967.

K. Popper. *The logic of scientific discovery*. Routledge, 2005.

D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

R. Précenth. Word embeddings and gender stereotypes in swedish and english, 2019.

A. Quinn and B. R. Quinn. *Figures of speech: 60 ways to turn a phrase*. Psychology Press, 1993.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

E. Reiter. 20 natural language generation. *The handbook of computational linguistics and natural language processing*, page 574, 2010.

E. Reiter and R. Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000. doi: 10.1017/CBO9780511519857.

M. Reuter. Swedish as a pluricentric language. *Pluricentric languages. Differing norms in different nations*, pages 101–116, 1992.

S. Rødven Eide, N. Tahmasebi, and L. Borin. The swedish culturomics gigaword corpus: A one billion word swedish reference dataset for nlp. 2016.

Y. Roh, G. Heo, and S. E. Whang. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2019.

S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL https://aclanthology.org/2021.eacl-main.24.

S. Rönnqvist, J. Kanerva, T. Salakoski, and F. Ginter. Is multilingual bert fluent in language generation? *arXiv preprint arXiv:1910.03806*, 2019.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, and C. Moldovan. Question generation shared task and evaluation challenge – status report. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 318–320, Nancy, France, Sept. 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2853.

S. S. Sabry, T. Adewumi, N. Abid, G. Kovacs, F. Liwicki, and M. Liwicki. Hat5: Hate language identification using text-to-text transfer transformer. *arXiv preprint arXiv:2202.05690*, 2022.

H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

M. Sammons, V. Vydiswaran, and D. Roth. Recognizing textual entailment. *Multilingual Natural Language Applications: From Theory to Practice*, pages 209–258, 2012.

P. Saxena and S. Paul. Epie dataset: A corpus for possible idiomatic expressions. In *International Conference on Text, Speech, and Dialogue*, pages 87–94. Springer, 2020.

A. P. Saygin and I. Cicekli. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258, 2002.

E. A. Schegloff. Sequencing in conversational openings 1. *American anthropologist*, 70 (6):1075–1095, 1968.

M. Schuster and K. Nakajima. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016. doi: 10.18653/v1/p16-1162.

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

S. M. Shieber. Lessons from a restricted turing test. *arXiv preprint cmp-lg/9404002*, 1994.

D. Shiffman, S. Fry, and Z. Marsh. *The nature of code*. D. Shiffman, 2012.

E. Shutova, D. Kiela, and J. Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1020. URL https://aclanthology.org/N16-1020.

A. Silvervarg, K. Raukola, M. Haake, and A. Gulz. The effect of visual gender on abuse in conversation with ecas. In *International conference on intelligent virtual agents*, pages 153–160. Springer, 2012.

J. Sim and C. C. Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.

A. Skrondal and S. Rabe-Hesketh. Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, 34(4):712–745, 2007.

E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL https://aclanthology.org/2020.acl-main.183.

N. A. Smith and J. Eisner. Annealing structural bias in multilingual weighted grammar induction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 569–576, 2006.

M. Snyder. When belief creates reality. In *Advances in experimental social psychology*, volume 18, pages 247–305. Elsevier, 1984.

I. Solaiman and C. Dennison. Process for adapting language models to society (palms) with values-targeted datasets. 2021. URL https://proceedings.neurips.cc/paper/2021/file/2e855f9489df0712b4bd8ea9e2848c5a-Paper.pdf.

C. Sporleder, L. Li, P. Gorinski, and X. Koch. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010a. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/618_Paper.pdf.

C. Sporleder, L. Li, P. Gorinski, and X. Koch. Idioms in context: The idix corpus. In *LREC*. Citeseer, 2010b.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

E. Stevens, L. Antiga, and T. Viehmann. *Deep Learning with PyTorch*. Manning, 2020.

Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020.

Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

H. Suresh and J. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9. 2021.

S. Traiger. Making the right identification in the turing test. In *The turing test*, pages 99–110. Springer, 2003.

J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

A. M. Turing et al. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5, 1936.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

V. Venekoski and J. Vankka. Finnish resources for evaluating language model semantics. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 231–236, 2017.

A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*, 2019.

P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8, 2019.

Z. Wang, A. W. Yu, O. Firat, and Y. Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021.

J. Weizenbaum. A computer program for the study of natural language. *Fonte: Stanford: http://web. stanford. edu/class/linguist238/p36*, 1969.

S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

M. West, R. Kraut, and H. Ei Chew. I'd blush if i could: closing gender divides in digital skills through education. 2019.

M. D. White. Immanuel kant. In *Handbook of economics and ethics*. Edward Elgar Publishing, 2009.

Wikipedia. English wikipedia multistream articles. 2019a. URL https://dumps.wikimedia.org/backup-index.html.

Wikipedia. Swedish wikipedia multistream articles. 2019b. URL https://dumps.wikimedia.org/backup-index.html.

Wikipedia. Wiki news abstract. 2019c. URL https://dumps.wikimedia.org/backup-index.html.

Wikipedia. Simple wiki articles. 2019d. URL https://dumps.wikimedia.org/backup-index.html.

Wikipedia. Yoruba wikipedia multistream articles, 2020. URL https://dumps.wikimedia.org/yowiki/20200801.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510, 2017.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

H. Youn, L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771, 2016.

L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018a.

S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018b. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020. doi: 10.48550/arXiv.1911.00536.

Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan. Joint retrieval and generation training for grounded text generation. *arXiv preprint arXiv:2105.06597*, 2021.

L. Zhou, J. Gao, D. Li, and H.-Y. Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020a.

M. Zhou, N. Duan, S. Liu, and H.-Y. Shum. Progress in neural nlp: Modeling, learning, and reasoning. *Engineering*, 6(3):275–290, 2020b. ISSN 2095-8099. doi: https://doi.org/10.1016/j.eng.2019.12.014. URL https://www.sciencedirect.com/science/article/pii/S2095809919304928.

L. Zhuang, L. Wayne, S. Ya, and Z. Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, Aug. 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.108.

Acronyms

AI Artificial Intelligence. 4, 10, 77, 102

ANN artificial neural network. ix, 1, 2, 13, 14, 23, 25, 27, 62

BART Bidirectional & Auto-Regressive Transformer. x, 87, 88

BERT Bidirectional Encoder Representations from Transformers. x, 37, 54, 56, 58, 87, 88

biLM bidirectional language model. 54

biLSTM bidirectional Long Short Term Memory Network. 46, 57, 58

BLEU bilingual evaluation understudy. 3, 21, 66, 70

BNC British National Corpus. 24, 31, 33, 34, 61, 71, 105

BoW bag-of-words. 41, 42, 45

BPE byte-pair encoding. 56, 67

BW Billion Word. 46, 48–50

CBoW continuous Bag-of-Words. 43, 44, 49–52

CC-BY4 Creative Commons Attribution 4.0. 11, 12, 100, 101

CI confidence interval. 20

CUS Credibility unanimous score. v, vi, 12, 26, 30, 72, 73, 95–97, 100, 106

DialoGPT Dialogue Generative Pre-trained Transformer. v, x, 24, 56, 68, 69, 71, 89–93, 95, 96, 99

ELMo Embeddings from Language Models. 53, 54, 56

GDC Gothenburg Dialogue Corpus. 62, 89, 90, 108

GDPR General Data Protection Regulation. 23

GMB Groningen Meaning Bank. 28, 46

GPT Generative Pre-trained Transformer. x, 24, 54, 66, 68

GUS Genial Understander System. 7

IAA Inter-Annotator Agreement. v, ix, 12, 26, 28–31, 57, 100, 104–106

IE Information Extraction. 63

IMDB Internet Movie Database. 28, 46

IR Information Retrieval. x, 5, 6, 9, 15, 21, 24, 39, 40, 62, 63, 66

LDA Latent Dirichlet Allocation. 42

LM language model. 53, 56, 68

LR learning rate. 53

LSI Latent Semantic Indexing. 42

LSTM Long Short Term Memory Network. 24, 46, 66, 89

MDL minimal dependency length. 80, 100

ML Machine Learning. v, 1, 27, 29, 36, 41, 69, 76, 100

MLM masked language model. 87, 88

MT Machine Translation. 6, 11, 15, 16, 25, 35, 36, 70, 87, 100, 102

MultiWOZ Multi-Domain Wizard-of-Oz. 35, 71, 106, 107

MWE Multi-Word Expression. 14–16, 40

NER Named Entity Recognition. x, 5, 22, 39, 40, 46, 47, 49, 51, 57, 63, 102

NLG Natural Language Generation. v, ix, x, 2, 3, 5, 7–9, 62, 63, 70

NLP Natural Language Processing. v, ix, 1, 2, 5, 6, 10, 11, 15, 21–23, 36, 42, 43, 56, 80, 100, 102, 106

NLTK natural language toolkit. 34, 40, 45

NLU Natural Language Understanding. v, 5, 6, 15, 62

NN neural network. ix, 13, 43, 45, 47, 49, 51, 56, 57, 65

OOV out-of-vocabulary. 43, 67

PCA Principal Component Analysis. 42

PCL Patronising and Condescending Language. 5

PIE Potential Idiomatic Expression. v, vi, ix, 11, 12, 16, 25, 28, 31, 33, 34, 36–38, 61, 71, 72, 99–101, 105, 109

PII personally identifiable information. 23, 73

PLSI Probabilistic Latent Semantic indexing. 42

PoS part of speech. 43

QA Question Answering. 6, 61

QG Question Generation. 6

RL reinforcement learning. 65, 66

RoBERTa Robustly optimized BERT pretraining Approach. 54

RQ research questions. 10–12, 100

RTE Recognizing Textual Entailment. 6

SA Sentiment Analysis. 5, 6, 46–49, 51

SoTA state-of-the-art. v, 20, 40, 49, 53, 54, 56, 67, 68, 80, 92, 99

SVD Singular Value Decomposition. 42

SVM support vector machine. 25

SW Simple Wiki. 46, 48–50

T5 Text-to-Text Transfer Transformer. x, 37, 38, 54, 67, 87

TC Text Classification. 5, 6

tf-idf term frequency-inverse document frequency. 40

UKWaC UK Web Pages. 31, 33, 34, 61, 71, 105

VS vector space. 41

VSM vector space model. 39, 41, 53

WSD Word Sense Disambiguation. 6, 15

XLM-R Cross-Lingual Model-RoBERTa. x, 57, 58, 87, 88

Department of SRT
Division of EISLAB

ISSN 1402-1544
ISBN 978-91-8048-055-0
ISBN 978-91-8048-056-7

Luleå University of Technology 2022

Print: Lenanders Grafiska, 5276632

LULEÅ
UNIVERSITY
OF TECHNOLOGY