Lexical Level Distribution of Metadiscourse in Spoken Language

Rui Correia
L²F, INESC-ID
Técnico Lisboa
Portugal
Rui.Correia@inesc-id.pt

Maxine Eskenazi
LTI
Carnegie Mellon University
USA
max@cs.cmu.edu

Nuno Mamede
L²F, INESC-ID
Técnico Lisboa
Portugal
Nuno.Mamede@inesc-id.pt

Abstract

This paper targets an understanding of how metadiscourse functions in spoken language. Starting from a metadiscourse taxonomy, a set of TED talks is annotated via crowdsourcing and then a lexical grade level predictor is used to map the distribution of the distinct discourse functions of the taxonomy across levels. The paper concludes showing how speakers use these functions in presentational settings.

1 Introduction

Often referred to as discourse about discourse, metadiscourse is the linguistic material intended to help the listener organize and evaluate the information in a presentation (Crismore et al., 1993). Examples include introducing (I’m going to talk about ...), concluding (In sum, ...), or emphasizing (The take home message is ...).

This paper explores how this phenomenon is used in spoken language, in particular how it occurs across presentations with different vocabulary levels. Are these acts used independently of vocabulary complexity? Which ones are used more frequently in more lexically demanding talks?

Finding out the answer to these questions has not only direct applications in language learning, but can also give insight on features that can be used for automatically classifying metadiscourse. Such classification establishes a link between discourse and lexical semantics, i.e., understanding the speaker’s explicit intention can be of help in tasks such as word sense disambiguation. For instance, the word means, in most contexts used to signal a definition, can also be used to show entailment, such as in: [...] these drugs [...] will reduce the number of complications, which means pneumonia and which means death.

This work was supported through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and the PhD Fellowship by FCT with reference SFRH/BD/51156/2010.

2 Related Work

The way discourse is used and organized in different grade levels started receiving attention in the early 80’s. Crismore (1980) focused on the use of a set of logical connectives at different levels and disciplines (high school through university), showing difficulty of mastery. McClure and Steffensen (1985) examined how linguistic complexity, developmental, and ethnic differences conditioned the use of conjunctions in children (3rd to 9th grade), finding a correlation between correct use of conjunctions and reading comprehension.

The first systematic approaches to metadiscourse were proposed by Williams (1981) and Meyer et al. (1980) and were further adapted and refined by Crismore (1983;1984) in a taxonomy that is still broadly used today. Crismore’s taxonomy is divided in two main categories: Informational and Attitudinal metadiscourse. The former deals with discourse and lexical semantics, i.e., understanding the speaker’s explicit intention can be of help in tasks such as word sense disambiguation. For instance, the word means, in most contexts used to signal a definition, can also be used to show entailment, such as in: [...] these drugs [...] will reduce the number of complications, which means pneumonia and which means death.

This paper is organized as follows. Section 2 presents related work on metadiscourse with focus to how it relates with grade level. Section 3 explains the choice of taxonomy of metadiscourse, describes the data and its annotation. Section 4 addresses the measure of vocabulary complexity used in this study, and the distribution of the data across different levels. Section 5 shows the results of mapping the metadiscourse functions according to vocabulary level. Finally, Section 6 has a discussion of the results and conclusions.
titude towards the discourse, and encompasses saliency (importance), emphatics (certainty degree), hedges (uncertainty degree), and evaluative (speaker attitude towards a fact).

Interestingly, it is in this early approach that we find the only attempt (to our knowledge) at understanding how metadiscourse occurs across grade levels. Crismore’s decisions while building the taxonomy are supported with examples extracted from nine social studies textbooks (elementary through college). After an annotation process, Crismore discusses the statistics and occurrence patterns of the various categories of metadiscourse across grade levels and audience. Goals were used very rarely in all text books. Pre-plans increased as students got into middle school and junior high and then declined. Post-plans were used when Pre-plans were used, about half as often. There was no clear trend toward increased use of Post-plans in upper grade texts. Topicalizers were used only at college level. Finally, for Attitudinal metadiscourse the author shows that it occurred more in texts which also contained more Informational metadiscourse, and that there was a tendency for it to increase in higher grade levels.

Intaraprawat and Steffensen (1995) also touched on the topic of metadiscourse and its relations to level, analyzing how 12 English as second language students used organizational language in their essays. When dividing them in good and poor, the authors observed that good essays contained proportionally more metadiscourse.

Regarding annotation of metadiscourse, and discourse in general, two distinct data-driven projects are broadly referred to and used. One is the Penn Discourse TreeBank (PDTB) (Weber and Joshi, 1998), built directly on top of Penn TreeBank (Marcus et al., 1993), composed of extracts from the Wall Street Journal. PDTB enriched the Penn TreeBank with discourse connectives annotation (conjunctions and adverbials), and organized them according to meaning (Mitksakaki et al., 2008). Given its goal to reach out to the NLP community and serve as training data, the resulting senses taxonomy is composed of low-level and fine-grained concepts.

In another approach, Marcu (2000) developed the RST Discourse Treebank, a semantics-free theoretical framework of discourse relations, intended to be “general enough to be applicable to naturally occurring texts and concise enough to facilitate an algorithmic approach to discourse analysis”. Similarly to PDTB, the RST Discourse Treebank is a discourse-annotated corpus intended to be used by the NLP community, based on Wall Street Journal articles extracted from the Penn Treebank. The difference between PDTB and the RST Discourse Treebank is the discourse organization framework, which in the case of the RST Discourse Treebank is the Rhetorical Structure Theory (Mann and Thompson, 1988).

All these approaches however, focus exclusively on written language. This was the motivation behind building our own corpora of metadiscourse in spoken language (see Section 3).

3 Metadiscourse Annotation

For this experiment we look at how metadiscourse is used in spoken English. We chose TED1, a source of self-contained presentations widely known for its speakers’ quality, and for targeting a general audience. A random sample of 180 talks was used, spanning several years and topics.

Our examination of theoretical underpinnings dealing with spoken language revealed that most approaches focus on the number of stakeholders involved, and never discuss function (Luukka, 1992; Mauranen, 2001; Auria, 2006). However, Ådel (2010) merges previous approaches in a taxonomy built upon MICUSP and MICASE (Römer and Swales, 2009; Simpson et al., 2002), corpora of academic papers and lectures, respectively.

Consequently, Ådel’s taxonomy was adapted according to the categories that appeared in the TED talks. More precisely, we consider 16 acts:

- **COM** – Commenting on Linguistic Form/meaning
- **CLAR** – Clarifying
- **DEF** – Definitions (originally Manage Terminology)
- **INTRO** – Introducing Topic
- **DELIM** – Delimiting Topic
- **CONC** – Concluding
- **ENUM** – Enumerating
- **POST** – Postponing Topic (originally Previewing)
- **ARG** – Arguing
- **ANT** – Anticipating Response
- **EMPH** – Emphasizing (originally Managing Message)
- **R & R** – collapse of Repairing with Reformulating
- **ADD** – collapse of Adding to Topic with Asides
- **EXMPL** – collapse of Exemplifying with Imagining Scenarios
- **RECAP** – Recapitulating (subdivision of the original Reviewing)
- **REFER** – Refer to Previous Idea (subdivision of the original Reviewing)

1https://www.ted.com/talks
Table 1: Annotation results in terms of occurrence (occur), confidence (conf) and agreement (α).

| Category | occur | conf | α  |
|----------|-------|------|----|
| ADD      | 93    | 3.88 | 0.15 |
| ANT      | 312   | 3.61 | 0.24 |
| ARG      | 283   | 3.51 | 0.32 |
| CLAR     | 265   | 3.82 | 0.15 |
| COM      | 203   | 3.10 | 0.33 |
| CONC     | 45    | 4.36 | 0.44 |
| DEF      | 169   | 4.04 | 0.29 |
| DELIM    | 26    | 4.21 | 0.31 |
| EMPH     | 330   | 3.31 | 0.18 |
| ENUM     | 343   | 3.74 | 0.49 |
| EXMPL    | 179   | 3.62 | 0.38 |
| INTRO    | 220   | 3.40 | 0.40 |
| POST     | 20    | 4.17 | 0.32 |
| RECAP    | 29    | 3.33 | 0.18 |
| REFER    | 76    | 3.93 | 0.32 |
| R&R      | 224   | 3.57 | 0.16 |

Crowdsourcing was used to annotate metadiscourse (Amazon Mechanical Turk\(^2\)). There was one task per metadiscursive category. This decreased the workers’ cognitive load per task. Each of the 180 talks was divided into segments of 500 words (truncated to the closest end of sentence). This configuration generated 742 Human Intelligent Tasks (HITs) for each category (not counting gold standard HITs). To annotate a given category, workers had to first pass a training session. Upon successful completion, they were asked to read each segment and select the words that signal the existence of the metadiscursive function in question. For agreement calculation and quality control purposes, each segment was annotated by 3 different workers.

Table 1 presents the annotation results, in terms of number of occurrences found (by majority vote), the average self-reported confidence on a 5-point Likert scale (1 equals to not confident at all and 5 equals to completely confident)\(^3\), and inter-annotator agreement. Herein, two workers are in agreement when the intersection of the words they select is not empty. A value of zero indicates complete disagreement, and α = 1 shows perfect agreement.

Results show that non-experts have trouble identifying some metadiscourse acts. Metadiscourse is a sparse phenomenon, even more so when dealt with one category at a time. It follows that the probability of two workers selecting the same passage by chance is very low. This quantity is taken into account when calculating agreement, and consequently, the case where one worker selects a word and others do not is severely penalized. Previous annotation attempts on similar phenomena, such as Wilson’s (2012) work on metalinguage, also show low agreement for sparser acts (0.09; 0.39), even when annotated by experts and considering only four categories.

Confidence results show all categories scoring above the middle of the scale (3). Workers showed less confidence for Commenting on Linguistic Form/Meaning (COM), which corresponds to the speaker commenting on their choice of words (confidence score of 3.1). On the other hand, workers showed the highest confidence for Concluding Topic (CONC), Delimiting Topic (DELIM) and Postponing Topic (POST), interestingly three categories that mark the change of topic in a talk.

4 Lexical Complexity

Evaluating linguistic complexity involves many aspects of language, such as lexis, syntax, semantics (Pilán et al., 2014; Dascalu, 2014). This paper, however, is concerned with the lexical complexity component only. Comparing the occurrences of metadiscourse across different vocabulary levels allows one to analyze its use independently of the syntactic structures that the speaker uses.

Although there is no commonly accepted measure of lexical complexity (Thériault, 2015), strategies typically rely on word unigrams to assure that only lexical clues are captured, since already capture grammatical properties (Vermeer, 2000; Heilman et al., 2007; Yasseri et al., 2011; Vajjala and Meurers, 2012). A drawback of such solutions is their inability of representing multi-word expressions, like fixed phrases or idioms.

This study uses the predictor described in Collins-Thompson and Callan (2004), which is available online\(^4\). This approach is a specialized Naive Bayes classifier with lexical unigram features only (for the previously mentioned reasons), which creates a model of the lexicon for each grade level - between 1\(^st\) and 12\(^th\).

\(^2\)https://www.mturk.com/mturk/welcome

\(^3\)Confidence measured on a sample of 100 segments

\(^4\)http://reap.cs.cmu.edu/demo/readability2012/
The training data is composed of 550 English documents evenly distributed across the 12 American grade levels, containing a total of 448,715 tokens and 17,928 types. The documents were drawn from a wide variety of subject areas such as fiction, non-fiction, history, science, etc.

All documents, comprised of both readings and student work, were collected online. Their level classification was directly extracted from the information contained in the web page that hosted them (for instance, a document extracted from a specific classroom page).

The system developed by Collins-Thompson and Callan (2004) first performs morphological stemming and stopword removal. Then, for a given passage $P$, the classifier computes the likelihood that the words of $P$ were generated from the representative language models of each level. The level where the likelihood is higher is the level that is attributed to $P$. The classifier performed at a correlation of 0.79 between the real and predicted levels (in a 10-fold cross validation setting).

It is important to note that this level prediction is used herein to distinguish between easier and more complex talks, more than to assign a specific grade level. In other words, one focuses at finding out which metadiscursive functions are used more often in talks with less demanding vocabulary with comparison to more complex ones (or vice-versa), never discussing occurrence at a specific level.

For the remainder of this study, the analysis takes place on two levels: whole talk and segment. The level predictor will be used on the 180 talks as a whole and on the 742 segments that compose them. The second strategy is a finer-grained local decision, since not all parts in a talk identified as high level are necessarily also complex.

Figure 1 shows level distribution: in black, the predictions when submitting the full talks to the classifier, and in light-gray, the segments prediction. In both cases we observe a normal distribution. It is interesting to notice the difference in the mode of the two cases. Most talks were assigned to a level corresponding to $8^{th}$ grade when submitted as a whole. However, when partitioned in segments, the most frequent level is the $7^{th}$.

To exclude the hypothesis that annotators’ performance was impacted by the complexity of the vocabulary, we examined how the vocabulary level of the talks relates with agreement.

Figure 2 shows how inter-annotator agreement is distributed. The correlation of the two variables is $\rho = 0.39$ for the talks and $\rho = 0.30$ for the segments, showing that vocabulary complexity does not negatively affect the capacity of two workers to agree on the annotation. In fact, the opposite trend was observed: workers agree more on segments with higher level vocabulary. This may be due to a higher degree of attention when facing more challenging content.

These results confirm that metadiscourse is independent of the content itself, and its structures can be detected independently of the propositional content in which they are inserted and for which they are used.

5 Results

With a set of 180 talks and 742 segments, annotated with 16 categories of metadiscourse, and automatically assigned to a level according to the lexical predictor described previously, one can now map the occurrences of the different acts across levels and conclude on how its use varies with lexical level of the content.
Table 2: Average occurrence and level correlation.

| Category          | Occur avg. (%) | Correlation by talk | Correlation by segment |
|-------------------|----------------|---------------------|------------------------|
| ADD               | 0.60           | 0.95                | 0.50                   |
| ANT               | 1.20           | (0.48)              | (0.85)                 |
| ARG               | 1.13           | 0.63                | 0.68                   |
| CLAR              | 1.47           | 0.58                | (0.16)                 |
| COM               | 1.54           | 0.78                | 0.70                   |
| CONC              | 0.37           | (0.07)              | (0.73)                 |
| DEF               | 1.13           | 0.63                | 0.85                   |
| DELIM             | 0.18           | 0.54                | 0.12                   |
| EMPH              | 1.90           | 0.47                | (0.27)                 |
| ENUM              | 3.15           | 0.09                | 0.23                   |
| EXMPL             | 1.47           | 0.43                | 0.50                   |
| INTRO             | 1.61           | 0.37                | 0.22                   |
| POST              | 0.21           | (0.21)              | (0.01)                 |
| RECAP             | 0.16           | 0.15                | (0.50)                 |
| REFER             | 0.54           | 0.94                | (0.33)                 |
| R&R               | 1.23           | 0.85                | 0.68                   |

Table 2 shows the probability of a sentence containing a given metadiscursive act (Occur avg. (\%)) and how each category correlates with level at both the talk and segment levels. Correlations are weighted for the amount of sentences in each level to decrease the impact of outliers in levels with few cases. Negative correlations are shown between brackets, significant correlations in bold, and high correlations are bold and underlined.

Adding Information (ADD) correlated at both talk and segment level, registering the highest correlation of all at talk level (0.95). Higher frequencies of ADD seem to be associated to talks with higher level vocabulary. This same pattern was also observed for R&R, which tends to occur in talks/segments assigned to higher grade levels.

Commenting on Linguistic Form/meaning (COM), and Definitions (DEF) also showed significant correlation at both levels. However, these categories have strong correlations at segment level, i.e., they do not only occur more frequently in higher level talks, but also in segments that contain words typically found in higher levels.

Anticipating Response (ANT) registered the strongest negative correlation both at sentence and talk levels (−0.85;−0.48). As talks are assigned to higher lexical levels, less instances addressing the audience’s previous knowledge are found. As one would expect, the more complex the vocabulary and topic of a talk is, the less assumptions are made about what the audience knows.

Arguing (ARG) shows moderate correlation at both levels. The more complex the vocabulary of a talk/segment is, the more the speaker feels the need to defend a point or prove his position.

Clarifications (CLAR) correlate moderately with the level of the talk but show a negative correlation trend with the segment. This shows that while talks with more demanding vocabulary have more clarifications, they are not necessarily located in lexically complex segments. This pattern is also observed for Conclusions (CONC), Recapitations (RECAP), and References to Previous Ideas (REFER), all with negative segment correlations. Interestingly, the four categories are related to paraphrasing (whether summarization or simplification). The high correlation for CONC in particular (0.73) shows that a segment that contains a conclusion tends to have simpler vocabulary.

Results for Delimiting Topic (DELM) and Exemplify (EXMPL) are at the frontier of low and moderate agreement. The remaining categories (Emphasizing, Enumerating, Introducing and Postponing Topic) did not correlate with level (\(\rho < 0.5\)) and seem to occur independently of the level of the vocabulary of the talk or segment.

6 Conclusions

This study used an empirical approach to understand how metadiscourse is used across different levels in spoken language. It employs a set of TED talks and a functional theory of metadiscourse. Crowdsourcing was used to annotate 16 metadiscourse functions. Comparing annotations with a vocabulary classifier showed that some but not all categories correlate with vocabulary level.

Strategies of topic management (delimiting, introducing, postponing) and broadly used functions (examples, emphasis, enumerations) occur at the same rate in all levels, not correlating with level.

Results also show that functions related to paraphrasing are more frequent in higher level talks, but not necessarily in segments containing the highest level vocabulary. In fact, the occurrence of a strategy that aims at language simplification contributes itself for lower level classification. This shift in correlation polarity from talk to segment level suggests that these strategies do not occur in close context with the ideas they are simplifying.

Contrastingly, functions that manage vocabulary (commentaries and definitions) seem to appear in the context of the vocabulary they address.

Future work includes using the annotation to build metadiscourse classifiers. As observed, the vocabulary level of the talk/segment can be a valuable feature for classification.
References

Annelie Ådel. 2010. Just to give you kind of a map of where we are going: A taxonomy of metadiscourse in spoken and written academic English. Nordic Journal of English Studies, 9(2):69–97.

Carmen PL Auria. 2006. Signaling speaker’s intentions: towards a phraseology of textual metadiscourse in academic lecturing. English as a Globalization Phenomenon. Observations from a Linguistic Microcosm, 3:59.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In HLT-NAACL, pages 193–200.

Avon Crismore, Raija Markkanen, and Margaret S Steffensen. 1993. Metadiscourse in persuasive writing a study of texts written by american and finnish university students. Written communication, 10(1):39–71.

Avon Crismore. 1980. Student use of selected formal logical connectors across school level and class type.

Mihai Dascalu. 2014. Analyzing discourse and text complexity for learning and collaborating. Studies in computational intelligence, 534.

Michael J Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In Proceedings of NAACL HLT, pages 460–467.

Puangpen Intaraprawat and Margaret S Steffensen. 1995. The use of metadiscourse in good and poor ESL essays. Journal of Second Language Writing, 4(3):253–272.

Klaus Krippendorff. 2007. Computing Krippendorff’s alpha reliability. Departmental Papers (ASC), page 43.

Minna-Riitta Luukka. 1992. Metadiscourse in academic texts. In Text and Talk in Professional Contexts. International Conference on Discourse and the Professions, Uppsala, 26-29 August, pages 77–88.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3):243–281.

Daniel Marcu. 2000. The Theory and Practice of Discourse Parsing and Summarization. The MIT press.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. Computational linguistics, 19(2):313–330.

Anna Mauranen. 2001. Reflexive academic talk: Observations from MICASE.

Erica F McClure and Margaret S Steffensen. 1985. A study of the use of conjunctions across grades and ethnic groups. Research in the Teaching of English, pages 217–236.

Bonnie JF Meyer, David M Brandt, and George J Bluth. 1980. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. Reading research quarterly, pages 72–103.

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the penn discourse treebank. In Computational Linguistics and Intelligent Text Processing, pages 275–286. Springer.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pages 174–184.

Ute Römer and John M. Swales. 2009. The michigan corpus of upper-level student papers (MICUSP). Journal of English for Academic Purposes, April.

Rita C Simpson, Sarah L Briggs, Janine Ovens, and John M Swales. 2002. The michigan corpus of academic spoken english. Ann Arbor, MI: The Regents of the University of Michigan.

Mélissa Thériault. 2015. The development of lexical complexity in sixth-grade intensive english students.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173. Association for Computational Linguistics.

Anne Vermeer. 2000. Coming to grips with lexical richness in spontaneous speech data. Language testing, 17(1):65–83.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In Coling/ACL workshop on discourse relations and discourse markers, pages 86–92.

Joseph M Williams. 1981. Ten lessons in clarity and grace. University of Chicago Press, Chicago.

T Yasseri, A Kornai, and J Kertész. 2011. A practical approach to language complexity: a Wikipedia case study. PloS one, 7(11):e48386–e48386.
Idiom Paraphrases: Seventh Heaven vs Cloud Nine

Maria Pershina    Yifan He    Ralph Grishman
Computer Science Department
New York University
New York, NY 10003, USA
\{pershina,yhe,grishman\}@cs.nyu.edu

Abstract
The goal of paraphrase identification is to decide whether two given text fragments have the same meaning. Of particular interest in this area is the identification of paraphrases among short texts, such as SMS and Twitter. In this paper, we present idiomatic expressions as a new domain for short-text paraphrase identification. We propose a technique, utilizing idiom definitions and continuous space word representations that performs competitively on a dataset of 1.4K annotated idiom paraphrase pairs, which we make publicly available for the research community.

1 Introduction

The task of paraphrase identification, i.e. finding alternative linguistic expressions of the same or similar meaning, attracted a great deal of attention in the research community in recent years (Bannard and Callison-Burch, 2005; Sekine, 2005; Socher et al., 2011; Guo et al., 2013; Xu et al., 2013; Wang et al., 2013; Zhang and Weld, 2013; Xu et al., 2015).

This task was extensively studied in Twitter data, where millions of user-generated tweets talk about the same topics and thus present a natural challenge to resolve redundancy in tweets for many applications, such as textual entailment (Zhao et al., 2014), text summarization (Lloret et al., 2008), first story detection (Petrovich, 2012), search (Zanzotto et al., 2011), question answering (Celikyilmaz, 2010), etc.

In this paper we explore a new domain for the task of paraphrase identification - idiomatic expressions, in which the goal is to determine whether two idioms convey the same idea.

This task is related to previous short-text paraphrase tasks, but it does not have access to many of the information sources that can be exploited in Twitter/short text paraphrasing: unlike tweets, idioms do not have hashtags, which are very strong topic indicators; unlike SMS, idioms do not have timestamp or geographical metadata; and unlike news headlines, there are no real world events that can serve as anchors to cluster similar expressions. In addition, an idea, or a moral of the idiom is often expressed in an indirect way, e.g. the idioms

(1) make a mountain out of a molehill
(2) tempest in a teapot

convey similar ideas:\(^{1}\):

(1) If somebody makes a mountain out of a molehill they exaggerate the importance or seriousness of a problem.
(2) If people exaggerate the seriousness of a situation or problem they are making a tempest in a teapot.

There is a line of research focused on extracting idioms from the text or identifying whether a particular expression is idiomatic (or a non-compositional multi-word expression) (Muzny and Zettlemoyer, 2013; Shutova et al., 2010; Li and Sporleder, 2009; Gedigian et al., 2006; Katz and Giesbrecht, 2006). Without linguistic sources such as Wiktionary, usingenglish.com, etc, it is often hard to understand what the meaning of a particular idiom is. It is even harder to determine whether two idioms convey the same idea or find alternative idiomatic expressions. Using idiom definitions, given by linguistic resources, one can view this problem as identifying paraphrases between definitions and thus deciding on paraphrases between corresponding idioms. Efficient techniques for identifying idiom paraphrases would complement any paraphrase identification system, and thus improve the downstream applications, such as question answering, summariza-

\(^{1}\)Definitions of these idioms are taken from http://www.usingenglish.com
tion, opinion mining, information extraction, and machine translation.

To the best of our knowledge we are the first to address the problem of determining whether two idioms convey the same idea, and to propose a new scheme that utilizes idiom definitions and continuous space word representation (word embedding) to solve it. By linking word- and sentence-level semantics our technique outperforms state-of-the-art paraphrasing approaches on a dataset of 1.4K annotated idiom pairs that we make publicly available.

2 Related Work

There is no strict definition of a paraphrase (Bhagat and Hovy, 2013) and in linguistic literature paraphrases are most often characterized by an approximate equivalence of meanings across sentences or phrases.

A growing body of research investigates ways of paraphrase detection in both supervised (Qiu et al., 2006; Wan et al., 2006; Das and Smith, 2009; Socher et al., 2011; Blacoe and Lapata, 2012; Madnani and Tetreault, 2012; Ji and Eisenstein, 2013) and unsupervised settings (Bannard and Callison-Burch, 2005; Mihalcea et al., 2006; Rus et al., 2008; Fernando and Stevenson, 2008; Islam and Inkpen, 2007; Hassan and Mihalcea, 2011). These methods mainly work on large scale news data. News data is very different from ours in two aspects: most news text can be interpreted literally and similar news events (passing a legislation, death of a person, elections) happen repeatedly. Therefore, lexical anchors or event anchors can work well on news text, but not necessarily on our task.

Millions of tweets generated by Twitter users every day provide plenty of paraphrase data for NLP research. An increasing interest in this problem led to the Paraphrase and Semantic Similarity In Twitter (PTI) task in SemEval-2015 competition (Xu et al., 2015). Existing bias towards Twitter paraphrases results in sophisticated systems that exploit character level similarity or metadata. But models relying on these insights are not necessarily applicable to other domains where misspellings are rare, or metadata is not available.

Idiomatic expressions constitute an essential part of modern English. They often behave idiosyncratically and are therefore a significant challenge for natural language processing systems. Recognizing when two idiomatic expressions convey similar ideas is crucial to recognizing the sentiment of the author, identifying correct triggers for events, and to translating the idiom properly. However, although there are several existing models to identify paraphrases in short text, idioms have very different characteristics from the data that those models are built on. In this paper, we experiment with two state-of-the-art paraphrasing models that are outperformed on our dataset of idiomatic expressions by a simple technique, raising a question on how well existing paraphrase models generalize to new data.

3 The Challenge

Identifying idiom paraphrases is an interesting and challenging problem. Lexical similarity is not a reliable clue to find similar idioms. Some idioms look very similar, differ in only one or two words, and convey the same idea. For example, “like two peas in a pod” vs “like peas in a pod” (“if people or things are like peas in a pod they look identical”), but other idioms that look similar can have very different meaning, e.g. “well oiled” vs “well oiled machine” (“if someone is well oiled they have drunk a lot” vs “something that functions very well is a well oiled machine”).

Finally, there are idioms that do not have any words in common at all and may seem quite different for a person not familiar with idiomatic expressions, but still have similar meaning. For example, “cross swords” vs “lock horns” (“when people cross swords they argue or dispute” vs “when people lock horns they argue or fight about something”). Thus, a natural way to identify idiom paraphrases is to focus on idiom definitions that explain meaning of an idiom in a clear and concise way.

4 Lexical vs Semantic Similarities

Our dataset consists of pairs \( \{\text{idiom}, \text{definition}\} \).

We use two types of similarity measures to compute how similar definitions of different idioms are: the lexical similarity is based on a lexical (word) overlap between two definitions, and the semantic similarity captures the overall semantic meaning of the whole sentence.

**Lexical similarity.** We compute cosine similarity between vectors \( \vec{v}_{d_1} \) and \( \vec{v}_{d_2} \), representing idiom descriptions \( d_1 \) and \( d_2 \) and weight each word
in these vectors by its tf-idf score:

\[ \text{lexSim}(d_1, d_2) = \cosine(\overrightarrow{v}_{d_1}, \overrightarrow{v}_{d_2}), \] (1)

where \( \overrightarrow{v}_d \) is a \(|V|\)-dimensional vector with \( V \) being the vocabulary of all definition words.

**Semantic similarity.** To capture the overall meaning of the definitions \( d \) we combine word embeddings (Collobert et al., 2011; Turian et al., 2010) for all words in \( d \) using two combination schemes:

- **Averaged sum:**
  \[
  \text{averaged}_d = \frac{1}{|d|} \sum_{\text{word} \in d} \text{emb}(\text{word})
  \] (2)

- **Weighted sum:**
  \[
  \text{weighted}_d = \frac{1}{\sum_{\text{word} \in d} \text{tfidf}_\text{word}} \sum_{\text{word} \in d} \text{tfidf}_\text{word} \cdot \text{emb}(\text{word})
  \] (3)

Then semantic similarity is measured as

\[ \text{semSim}(d_1, d_2) = \cosine(\overrightarrow{\text{comb}}_{d_1}, \overrightarrow{\text{comb}}_{d_2}) \] (4)

where \( \overrightarrow{\text{comb}}_d \) is a 100-dimensional vector combined from word embeddings \( \text{emb}(\text{word}) \) (Turian et al., 2010) for words in description \( d \) using either averaged (2) or weighted (3) combination schemes.

4.1 IdiomSim

There is a tradeoff between the two similarity measures \( \text{lexSim} \) and \( \text{semSim} \) (Section 4); while the first one captures the actual lexical overlap, the second one can better capture the closeness in semantic meaning. To find an optimal balance between the two we consider their weighted sum

\[ \text{IdiomSim}(d_1, d_2) = (1 - \alpha) \cdot \text{lexSim}(d_1, d_2) + \alpha \cdot \text{semSim}(d_1, d_2) \] (5)

and decide on \( \alpha \) by optimizing for a maximal F-score on a development dataset.

5 Experiments

**Data.** We collected 2,432 idioms from http://www.usingenglish.com, a site for English learners, where every idiom has a unique description giving a clear explanation of the idiom’s meaning. As opposed to tweets there are no hashtags, no topics or trends, no timestamps, or any other default evidence, that two idioms may convey similar ideas. Thus it becomes a challenging task itself to construct a dataset of pairs that is guaranteed to have a certain fraction of true paraphrases.

We used a simple cosine similarity between all possible idiom definitions pairs to have a ranked list and labeled the top 1.5K pairs. Three annotators were asked to label each pair of idiom definitions as “similar” (score 2), “have something in common” (score 1), “not similar” (score 0). 0.1K pairs received a total score of 4 (either 2+2+0, or 2+1+1), and were further removed as debatable. The rest of the labeled pairs were randomly split into 1K for test data and 0.4K for development. Only pairs that received a total score of 5 or higher were considered as positive examples. There are 364 and 96 true paraphrases in our test and development sets respectively. ²

**Baselines.** Our baselines are simple and tf-idf weighted cosine similarity between idiom description sentences: CosSim and LexicalSim.

We compare our method with the deterministic state-of-the-art ASOBEK model (Eyecioglu and

²We use 100-dimensional Turian word embeddings available at http://metaoptimize.com/projects/wordreprs/
³https://github.com/masha-p/Idiom_ Paraphrases
| Idioms                        | Descriptions                                                                 |
|------------------------------|-----------------------------------------------------------------------------|
| seventh heaven               | if you are in seventh heaven you are extremely happy                       |
| cloud nine                   | if you are on cloud nine you are extremely happy                            |
| face only a mother could love| when someone has a face only a mother could love they are ugly              |
| stop a clock                 | a face that could stop a clock is very ugly indeed                          |
| take your medicine           | if you take your medicine you accept the consequences of something          |
| face the music               | you have done wrong                                                         |
| well oiled                   | if someone is well oiled they have drunk a lot                              |
| drunk as a lord              | someone who is very drunk is as drunk as a lord                             |
| cheap as chips               | if something is very inexpensive it is as cheap as chips                    |
| to be dog cheap              | if something is dog cheap it is very cheap indeed                           |
| great minds think alike      | if two people have the same thought at the same time                        |
| on the same wavelength       | if people are on the same wavelength they have the same ideas and opinions about something |
| could eat a horse            | if you are very hungry you could eat a horse                                |
| hungry as a bear             | if you are hungry as a bear it means that you are really hungry             |
| cross swords                 | when people cross swords they argue or dispute                              |
| lock horns                   | when people lock horns they argue or fight about something                 |
| talk the hind legs off a donkey| a person who is excessively or extremely talkative                          |
| talk the legs off an iron pot| can talk the hind legs off a donkey                                         |
|                             | somebody who is excessively talkative or is especially convincing           |
|                             | is said to talk the legs off an iron pot                                    |

Table 1: Examples of extracted idiom paraphrases.

Keller, 2015) that was ranked first among 19 teams in the Paraphrase in Twitter (PIT) track on the SemEval 2015 shared task (Xu et al., 2015). This model extracts eight simple and elegant character and word features from two sentences to train an SVM with linear kernel. It achieves an F-score of 55.1% on our test set. 4

We also compare our method with the state-of-the-art Weighted Textual Matrix Factorization model (WTMF) (Guo et al., 2013), 5 which is specifically developed for short sentences by modeling the semantic space of words, that can be either present or absent from the sentences (Guo and Diab, 2012). This model achieves a maximal F-score of 61.4% on the test set.

The state-of-the-art model for lexically divergent paraphrases on Twitter (Xu et al., 2015) is tailored for tweets and requires topic and anchor words to be present in the sentence, which is not applicable to idiom definitions.

Evaluation and Results. To evaluate models we plot precision-recall curves for CosSim, WTMF, LexicalSim, and IdiomSim (for clarity we omit curves for other models). We also compare maximal F-score for all models. We observe that simple cosine similarity (CosSim) achieves a maximal F-score of 53.7%, LexicalSim is a high baseline and achieves an F-score of 63.75%. When we add averaged word embeddings the maximal F-score is 64.4% (IdiomSim ave). With tfidf weighted word embeddings we achieve F-score of 65.9% (IdiomSim +). By filtering out uninformative words such as “a”, “the”, etc (12 words total) we improve the F-score to 66.6% (IdiomSim +). With tfidf weighted word embeddings we achieve F-score of 65.9% (IdiomSim +), outperforming state-of-the-art paraphrase models by more than 5% absolute (Figure 1). Both IdiomSim and IdiomSim + outperform WTMF significantly according to a paired t-test with p less than 0.05.

Examples and Discussion. We use threshold, corresponding to a maximal F-score obtained on the development dataset, and explore paraphrases from test dataset scored higher and lower than this threshold. Examples of extracted idiom paraphrases are in Table 1. Examples of false positives and false negatives are in Table 2.

Simple word overlap is not a reliable clue to de-
## Idioms

| False positives               | Descriptions                                                                 |
|------------------------------|------------------------------------------------------------------------------|
| healthy as a horse           | if you are as healthy as a horse you are very healthy                        |
| an apple a day keeps         | if you are as healthy as a horse you are very healthy                        |
| the doctor away              | eating healthy food keeps you healthy                                        |
| jersey justice               | jersey justice is a very severe justice                                      |
| justice is blind             | justice is blind means that justice is impartial and objective              |
| heart of steel               | when someone has a heart of steel they do not show emotion or are not affected emotionally |
| heart of glass               | when someone has a heart of glass they are easily affected emotionally       |

### False negatives

| Example                      | Description                                                                 |
|------------------------------|------------------------------------------------------------------------------|
| like a kid in a candy store  | if someone is like a kid in a candy store                                 |
| bee in your bonnet           | if someone is very excited about something                                |
| easy as falling off a log    | something very easy or simple to do is as easy as falling off a log        |
| no sweat                     | no sweat means something is easy                                           |
| hopping mad                  | if you are hopping mad you are extremely angry                              |
| off on one                   | if someone goes off on one they get extremely angry indeed                 |

Table 2: Examples of false positive and false negative paraphrases.

Our future work will be focused on exploring different strategies to compute semantic similarity between sentences, developing a comprehensive idiom similarity measure that will utilize both idioms and their definitions, and on comparing text with an idiom and a general text as a realistic scenario for paraphrase identification. It is a new and a challenging task and thus opens up many opportunities for further research in paraphrase identification and all its downstream applications.

### Acknowledgments

We thank Thien Huu Nguyen of New York University and Asli Eyecioglu of University of Sussex for their help and advice.

### References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. (2012). SemEval-2012 task 6: A pilot on semantic textual similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM*).

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larrainz Uria, and Janice Wiebe. (2015). SemEval-2015 task 2: Semantic textual similarity, English, Spanish and Pilot on Interpretability. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).
Weiwei Guo and Chris Callison-Burch. (2005). Paraphrasing with Bilingual Parallel Corpora. In Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL).

Marco Baroni, Georgiana Dinu, and German Kruszewski. (2014). Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL).

Rahul Bhagat and Eduard Hovy. (2013). What is a paraphrase? In Proceedings of the International Conference on Computational Linguistics (COLING).

Julia Birke and Anoop Sarkar. (2006). A clustering approach for nearly unsupervised recognition of non-literal language. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL).

William Blacoe and Mirella Lapata. (2012). A comparison of vector-based representations for semantic composition. In Proceedings of EMNLP-CoLNL.

Asli Celikyilmaz, Dilek Hakkani-Tur, and Gokhan Tur. (2010). LDA based similarity modeling for question answering. In Proceedings of the NAACL HLT 2010 Workshop on Semantic Search.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). Natural Language Processing (Almost) from Scratch. In Journal of Machine Learning Research (JMLR).

Dipanjan Das and Noah A. Smith. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language (ACL-IJCNLP).

Asli Eyecioglu and Bill Keller. (2015). ASOBEK: Twitter Paraphrase Identification with Simple Overlap Features and SVMs In Proceedings of 9th International Workshop on Semantic Evaluation (SemEval).

Samuel Fernando and Mark Stevenson. (2008). A semantic similarity approach to paraphrase detection. Computational Linguistics UK (CLUK) 11th Annual Research Colloquium.

Matt Gedigian, John Bryant, Srin Narayanan, and Branimir Ciric. (2006). Catching metaphors. In Proceedings of the Third Workshop on Scalable Natural Language Understanding (ScaNaLU).

Weiwei Guo and Mona Diab. (2013). Modeling Sentences in the Latent Space. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL).

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. (2013). Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL).

Samer Hassan and Rada Mihalcea. (2011). Semantic relatedness using salient semantic analysis. In Proceedings of the twenty-fifth Association for the Advancement of Artificial Intelligence Conference (AAAI).

Aminul Islam and Diana Inkpen. (2007). Semantic similarity of short texts. In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP).

Yangfeng Ji and Jacob Eisenstein. (2013). Discriminative improvements to distributional sentence similarity. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Graham Katz and Eugenie Gireshbrecht. (2006). Automatic identification of non-compositional multiword expressions using latent semantic analysis. In Proceedings of the Workshop on Multitword Expressions: Identifying and Exploiting Underlying Properties (MWE).

Linlin Li and Caroline Sporleder. (2009). Classifier combination for contextual idiom detection without labeled data. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Elena Lloret, Oscar Ferrandez, Rafael Munoz, and Manuel Palomar. (2008). A text summarization approach under the influence of textual entailment. In Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS).

Nitin Madnani and Joel Tetreault. (2012). Re-examining machine translation metrics for paraphrase identification. In Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT).

Rada Mihalcea, Courtney Corley, and Strapparava. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference (AAAI).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffery Dean. (2013). Efficient estimation of word representations in vector space. In Proceedings of Workshop at the International Conference on Learning Representations (ICLR).

Grace Muzny and Luke Zettlemoyer. (2013). Automatic Idiom Identification in Wiktionary. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP).
Sasa Petrovic, Miles Osborne, and Victor Lavrenko (2012). Using paraphrases for improving first story detection in news and Twitter. In Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT).

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. (2006). Paraphrase recognition via dissimilarity significance classification. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP).

Vasile Rus, Philip M. McCarthy, Mihai C. Linteau, Danielle S. McNamara, and Arthur C. Graesser (2008). Paraphrase identification with lexicosyntactic graph subsumption. In Proceedings of the Twenty-First International FLAIRS Conference.

Satoshi Sekine. (2005). Automatic paraphrase discovery based on context and keywords between NE pairs. In Proceedings of the 3rd International Workshop on Paraphrasing.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. (2002). Automatic paraphrase acquisition from news articles. In Proceedings of the 2nd International Conference on Human Language Technology Research (HLT).

Ekaterina Shutova, Lin Sun, and Anna Korhonen. (2010). Metaphor identification using verb and noun clustering. In Proceedings of the International Conference on Computational Linguistics (COLING).

Richard Socher, Eric H Huang, Jeffrey Pennington, Andrew Y Ng, and Christopher D Manning. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In Proceedings of Advances in Neural Information Processing Systems (NIPS).

Joseph Turian, Lev Ratinov, and Yoshua Bengio. (2010). Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL).

Stephen Wan, Mark Dras, Robert Dale, and Cecile Paris. (2006). Using dependency-based features to take the paraphrase out of paraphrase. In Proceedings of the Australasian Language Technology Workshop.

Ling Wang, Chris Dyer, Alan W Black, and Isabel Trancoso. (2013). Paraphrasing 4 microblog normalization. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP).

Wei Xu, Chris Callison-Burch, and William B. Dolan. (2015). Extracting Lexically Divergent Paraphrases from Twitter. Transactions of the Association for Computational Linguistics (TACL).

Wei Xu, Alan Ritter, and Ralph Grishman. (2013). Gathering and Generating Paraphrases from Twitter with Application to Normalization. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora (BUCC).

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Kostas Tsioutsiouliklis (2011). Linguistic redundancy in Twitter. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP).

Congle Zhang and Daniel S. Weld (2013). Harvesting parallel news streams to generate paraphrases of event relations. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP).

Jiang Zhao, Man Lan, Zheng-Yu Niu, and Dong-Hong Ji. (2014). Recognizing cross-lingual textual entailment with co-training using similarity and difference views. In Proceedings of International Joint Conference on Neural Networks (IJCNN).

Jiang Zhao, Man Lan, and Jun Feng Tian. (2015). ECNU: Using Traditional Similarity Measurements and Word Embedding for Semantic Textual Similarity Estimation. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).