Literal or idiomatic? Identifying the reading of single occurrences of German multiword expressions using word embeddings

Rafael Ehren
Dept. of Computational Linguistics
Heinrich Heine University
Düsseldorf, Germany
Rafael.Ehren@hhu.de

Abstract

Non-compositional multiword expressions (MWEs) still pose serious issues for a variety of natural language processing tasks and their ubiquity makes it impossible to get around methods which automatically identify these kind of MWEs. The method presented in this paper was inspired by Sporleder and Li (2009) and is able to discriminate between the literal and non-literal use of an MWE in an unsupervised way. It is based on the assumption that words in a text form cohesive units. If the cohesion of these units is weakened by an expression, it is classified as literal, and otherwise as idiomatic. While Sporleder an Li used *Normalized Google Distance* to model semantic similarity, the present work examines the use of a variety of different word embeddings.

1 Introduction

Non-compositional multiword expressions (MWEs) still pose serious issues for a variety of natural language processing (NLP) tasks. For instance, if you use the free machine translation service Google Translate to translate example[1] (1-a) from English to German, according to the translation (1-b) the stabbing (luckily for John) doesn't cause his immediate death, but him literally kicking a bucket.

(1) a. Because John was stabbed, he kicked the bucket.
'Because John was stabbed, he died.'

b. Weil John erstochen wurde, trat er den Eimer.
'Because John was stabbed, he stroke a pail with his foot.'

Although not an absolutely impossible scenario, the context strongly suggests that *kicked the bucket* is not meant literally in (1-a) and therefore a literal translation is not the desired one.

Such errors illustrate the necessity for methods which automatically identify occurences of idiomatic MWEs when there is also a literal counterpart. Thus, there are actually two different identification tasks:

1. Determine wheter an MWE can have an idiomatic meaning;

2. Determine which of the two possible meanings, namely the literal and the idiomatic one, an MWE has given a specific context.

For example (1-a) this would mean to first figure out whether *kick the bucket* has another meaning than 'to strike a pail with one's foot' and then to decide which meaning it has in the context of the sentence. This paper is mainly concerned with the second task.

The method presented in this paper was inspired by the work of Sporleder and Li (2009) and is based on the assumption that words and sentences in a text are not completely independent of each other regarding their meaning, but form topical units. This relatedness between words is termed *lexical cohesion*. Sequences of words which exhibit a cohesive relationship are called *lexical chains* (Morris and Hirst, 1991). The intuition behind the approach is that idioms weaken this cohesion, because they often contain elements that are used in a figurative sense and thus do not "fit" into their contexts. If, for example, the MWE

[1] All of the examples presented in this paper were invented by the author.

Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 103–112, Valencia, Spain, April 3-7 2017.

break the ice is used in a literal sense, it will very likely co-occur with terms that are topically related like *snow, water, iceberg, etc.* This is usually not the case for the idiomatic use of *break the ice*. Consider the following example:

(2) For his future bride's sake he wanted to break the ice between him and his prospective parents-in-law before the wedding.

In (2), the expression *ice* appears with words (*wife*, *parents-in-law*, *wedding*) that do not belong to the same topical field as the literal meaning of *ice* and therefore it is not part of the dominating lexical chain.

Sporleder and Li made use of this fact and built cohesion-based classifiers to automatically distinguish between the literal and idiomatic version of an MWE. Following Sporleder and Li, we also implemented a classifier based on textual cohesion, albeit using a different measure for semantic similarity. While Sporleder and Li relied on *Normalized Google Distance* (NGD), a measure that uses the number of results for a search term as a basis, different word embeddings[2] were used in the context of this work. Word embeddings seemed like a more promising way of representing the meaning of words since a plain co-occurence-based approach like the NGD has some considerable limitations as we will discuss in section 3.2. Furthermore, a comparison of different types of embeddings was conducted where it became apparent that the implemented vector spaces are not all equally well suited for the task at hand. The task was conducted with a total of three different vector spaces and some achieved better results than others. Finally the best performing vector space was used to compare the effect of different window sizes around the MWE.

2 Related Work

Hirst and St-Onge (1998) followed the notion that words in a text are cohesively tied together and used it to detect and correct malapropisms. A malapropism is the erroneous use of a word instead of a similar sounding word, caused by a typing error or ignorance of the correct spelling. For instance: *It's not there fault*. In this sentence the adverb *there* is mistakenly used in place of the possessive determiner *their*. Since they are correctly spelled, malapropisms cannot be detected by spelling checkers that only check the orthography of a word. To tackle this problem, Hirst and St-Onge represented context as lexical chains and compared the words that did not fit into these chains with orthographically similar words. Semantic similarity was determined using WordNet.

Sporleder and Li (2009) were inspired by Hirst and St-Onge's method and applied it to MWEs, which they treated analogously to malapropisms. In their experiments the idiomatic version of an MWE is equivalent to a malapropism, because it usually does not participate in the lexical chains constituting the topic(s) of a text. Accordingly the literal sense of an MWE would be the correct word if we stay within the analogy. However, in contrast to Hirst and St-Onge, they did not rely on a thesaurus to model semantic similarity, but on NGD. As already stated in the introduction, NGD is a measure for semantic similarity that uses the number of pages returned by a search engine as a basis and is calculated as follows:

$$NGD(x,y) = \frac{max\{logf(x),logf(y)\} - logf(x,y)}{logN - min\{logf(x),logf(y)\}}$$

The number of pages for the search terms *x* and *y* are given by *f(x)* and *f(y)*, the number of pages containing *x AND y* by *f(x,y)*. *N* denotes the total number of web pages indexed by the search engine. If we take a look at the numerator we can see that it gets smaller the more often the two terms occur together. So an NGD of *0* means *x* and *y* are as similar as possible, while they get a score of greater or equal to *1* if they are very dissimilar.

With the NGD as a measure of semantic similarity, Sporleder and Li implementend two unsupervised cohesion-based classifiers that had the task to discriminate between the literal and non-literal use of an MWE. One of these classifiers did this based on the question whether a given MWE participated in one of the lexical chains in a text. If it did, the MWE was labeled as literal, if not, as idiomatic. The other classifier built *cohesion graphs* and made this decision based on whether the graph changed when the expression was part of the graph or left out (cohesion graphs will be elucidated in section 3.3).

Katz and Giesbrecht (2006) also examined a method to automatically decide whether a given MWE is used literally or idiomatically. Their method relied on word embeddings which were

[2]Word embedding is a collective term to denote the mapping of a word to a vector.

obtained through *Latent Semantic Analysis* (LSA). The experiment was conducted as follows: In a first step, Katz and Giesbrecht annoted for 67 instances of the German MWE *ins Wasser fallen* according to whether they were used literally or non-literally in their respective context.[3] Subsequently they generated a vector for the literal and a vector for the idiomatic use of the expression. In order to determine the meaning of the MWE with regard to the context, a nearest-neighbour classification was performed.

3 Setup

3.1 Lexical Cohesion

The term cohesion describes the property of a text that its items are not independent from one another, but somehow "tied together". Cohesion manifests itself in three different ways: back-reference, conjunction and semantic word relations (Morris and Hirst, 1991). Back-reference is usually realised through the use of pronouns (***Sarah*** *went to the dentist.* ***She*** *had a toothache.*). Conjunctions link clauses together and explicitly interrelate them (*John went home,* ***because*** *he was drunk*). But the only manifestation of cohesion significant for the present work are the semantic relations between the words in a text, i.e. the lexical cohesion. Lexical cohesion can be divided into five classes (Morris and Hirst, 1991; Stokes et al., 2004):

1. Repetition: *Kaori went into the* ***room****. The* ***room*** *was dark.*

2. Repetition through synonymy: *After a short rest Sally mounted her* ***steed****. But the* ***horse*** *was just too tired to go on.*

3. Repetition through specification/generalisation: *Shortly after he ate the* ***fruit****, his stomach began to cramp badly. It seemed that the* ***apple*** *was poisoned.*

4. Word association through a systematic semantic relationship (e.g. meronymy): *The* ***team*** *seemed unbeatable at that time. Already when the* ***players*** *went out on court, they put the fear of god in their opponents.*

5. Word association through a nonsystematic semantic relationship: *The* ***party*** *started at sunset. They* ***danced*** *till sunrise.*

[3]The literal meaning is 'to fall into the water', the idiomatic meaning is 'to fail to happen'.

Semantic relations like antonymy (*quiet*, *loud*), hyponymy (*bird*, *sparrow*) or meronymy (*car*, *tire*) are classified under systematic relationships. However, it is not always possible to specify the systematics behind a relationsip holding between two words (*party*, *to dance*). But for our purpose, it is not really necessary to identify the exact semantic relation, one only has to recognize that there is one. Even if we can't state what relation holds between *party* and *to dance*, we know that they are topically, and thus semantically, close.

Sequences of words exhibiting the forms of lexical cohesion listed above are referred to as lexical chains. These sequences, which can be more than two words long and cross sentence boundaries, span the topical units in a text (Morris and Hirst, 1991). In other words, they indicate what a text is about. That is why lexical chains can play an important role in text segmentation and summarization. The following example shows such a cohesive chain:

(3) When the ice finally broke the ice bear jumped off his floe into the ocean and fled. The icebreaker was designed to cut through the thickest ice, but soon it showed that even this huge ship could not withstand the unforgiving cold of the arctic. They had backed the wrong horse.

If we consider only the nouns in example (3) a possible lexical chain would be *ice*, *ice bear*, *floe*, *ocean*, *icebreaker*, *ice*, *ship*, *cold*, *arctic*. It indicates that the text segment is about the act of breaking sea ice. The lexical cohesion shows itself by repetition through generalisation (*icebreaker*, *ship*), repitition (*ice*, *ice*) and word association through unsystematic semantic relationships (e.g. *cold*, *arctic*). The only noun arguably not linked to any of the other words by a semantic relation and hence not participating in the cohesive chain is *horse*, the noun component of the idiomatic expresion *to back the wrong horse*. One could maybe argue that *horse* and *ice bear* share some semantic content since they are both four-legged mammals, but apart from that the case is pretty clear: *horse* is not part of the topical unit which is about the act of breaking sea ice. Therefore it's possible to conclude that *back the wrong horse* is not meant literally in this context.

Thus by looking for missing cohesive links one

is able to detect idiomatic readings of MWEs. In order to automatize this process, it is necessary to measure the semantic relatedness of two words. And to do that, it is in turn necessary to first model the meaning of words.

3.2 Word Embeddings

For their experiments Sporleder and Li (2009) modelled the semantic similarity of words in terms of the NGD. The advantage of the NGD is that no corpus can compare in size and up-to-dateness to the (indexed) web, which means that information regarding the words one is looking for is very likely to be found (Sporleder and Li, 2009).

Nevertheless, the method has some drawbacks. As Sporleder and Li state themselves, the returned page counts for the search terms can be somewhat unstable which is why they used Yahoo to obtain the web counts instead of Google because the former delivered more stable counts. Furthermore they had to leave out very high frequency terms because neither the Google nor the Yahoo API would deliver reliable results for those. But these are only minor issues compared to the fact that NGD is not the most sophisticated way of representing the semantics of words. The NGD reduces semantic similarity to the question of how often two terms occur together in a specific context relative to their total frequency. Although this simplification works surprisingly well, we will see herinafter that it has its limitations.

The basis for the representation of word meaning with distributional patterns is the distributional hypothesis. It states that words that occur in similar contexts have similar meanings. Or as John Rupert Firth prominently phrased it:

> "You shall know a word by the company it keeps!" (Firth, 1957, p. 11)

As an example, Firth gives the term *ass* which, according to him, is in familiar company with phrases like *you silly...*, *he is a silly...* or *don't be such an...* Not only would English speakers be able to guess with a certain probability which term they had to fill in for the dots, but other guesses presumably would fall on semantically similar words like *jerk*, *fool* or *idiot*. The validity of the distributional hypothesis and the fact that people only need a very small context window to infer the meaning of a word has been shown in different experiments (Rubenstein and Goodenough, 1965; Miller and Charles, 1991).

From the distributional hypothesis one can conclude that the semantic similarity of words does not manifest itself only through co-occurence (as the NGD simplifies), but also through shared neighbourhood. It might even be the case that some semantically very similar words appear less often together than one would expect, for example if a synonym is used to the exclusion of the other. Sahlgren (2006) did an experiment which strengthens this suspicion. He created two different representations of word meaning in form of vector spaces[4], one with a syntagmatic use of context and one with a paradigmatic use of context[5]. Then Sahlgren conducted the TOEFL synonym test[6] with both vector spaces and found that the paradigmatic word space achieved better results (75%) than the syntagmatic word space (67.5%). Sahlgren furthermore states that LSA performed on word-document matrices increases the results of TOEFL experiments because it reveals the "hidden" concepts behind words and thus relates words which do not co-occur, but appear in similar documents. This way, according to Sahlgren, a paradigmatic use of contexts is approximated. This shows that methods relying only on the co-occurence of words (syntagmatic relations) like the NGD are limited when it comes to the representation of word meaning. For that reason it seems more promising to model semantic relatedness with word embeddings, specifically word embeddings that represent syntagmatic **and** paradigmatic relations between words.

Word embeddings that incorporate a paradigmatic use of context by design are those who originate from the construction of a word-context matrix. But like documents in a term-document matrix, words in the word-context matrix are still only represented by bag-of-words. That is why structural vector space models (VSM) of word meaning were developed. These models, as one can already guess from the name, contain structural information about the words in the corpus,

[4]Words were represented by context vectors, NGD was not used in the experiment. But as it is the case with NGD one of the representations was created only considering co-occurence counts in a specific context region.

[5]A syntagmatic relation holds between to words that co-occur together, a paradigmatic relation holds between to words that share neighbours (i.e. they are potentially interchangeble).

[6]The TOEFL synonym test is a test were the testee has to choose the correct synonym for a given word out of four candidates (e.g. target word: levied; candidates: imposed, believed, requested, correlated; correct answer: imposed).

e.g. grammatical dependencies (Padó and Lapata, 2007). A model enriched with such information would, for example, be able to capture the fact that *the dog* is the subject and does the biting in the sentence *the dog bites the man*. A dimension of the word *dog* could thus be *sbj_intr_man*. The hope is that these models do a better job at representing semantics, because they take word order into account and ensure that there is an actual lexico-syntactic relation between the target and the context word and not only a co-occurence relationship.

An alternative to the "classic" count-based approach for the creation of word embeddings are skip-gram and continuous bag-of-words (CBOW). Skip-gram and CBOW, often grouped under the term *word2vec*, are two shallow neural networks which are able to create low-dimensional word embeddings from very large amounts of data in a relatively short amount of time. These two properties paired with the fact that the resulting word representations perform really well explain why word2vec has gained a lot of traction since Mikolov et al. (2013a; 2013b) presented it in 2013. In contrast to the "common" way of creating word embeddings by first constructing a word-context matrix of high dimensionality and then reducing the dimensions with LSA, word2vec creates low-dimensional vectors right from the start. This is possible, because skip-gram and CBOW do not count co-occurences in the corpus, but try to predict words. The skip-gram model tries to predict the neighbours of a word *w*, while CBOW tries to predict *w* from its neighbours. The intuition behind this approach is that a representation of a word that is good at predicting its surrounding words is also a good semantic represenation since words in similar contexts tend to have similar meanings (Baroni et al., 2014).

Levy et al. (2015) succeeded in showing that the perceived superiority of word2vec over traditional count-based methods (Baroni et al., 2014) is not founded in the algorithms themselves, but in the choice of certain parameters (Levy et al. call them "hyperparameters") which can be transferred to traditional models. Furthermore they showed that skip-gram with negative sampling (SGNS) implicitly generates a word-context matrix whose elements are Pointwise Mutual Information (PMI)[7] values shifted by a global constant. Hence, the data basis for word2vec and for the conventional methods is maybe not that different after all.

3.3 Experimental setup

To disambiguate between the literal and non-literal meaning of German MWEs it was of course necessary to first find instances of such MWEs. Those instances (along with the containing paragraphs) were automatically extracted from the TüPP-D/Z (Tübinger partiell geparstes Korpus - Deutsch/Zeitung)[8] corpus, a collection of articles from the German newspaper *die tageszeitung* (taz) from the years 1986 – 1999. Then the instances were annotated by hand depending on whether their readings were literal or idiomatic.

The MWEs listed in table 1 were chosen, because they are a part of figurative language and have a literal and idiomatic meaning. The latter is not self-evident, since some figurative MWEs do not have a literal meaning due to their syntactic idiosyncrasy, e.g. *kingdom come* and *to trip the light fantastic*.[9] And even the ones who do are mostly used in an idiomatic sense as one can see from the total count in table 1. 85% of the instances were used idiomatically.

MWE	Literal	Idiomatic	Total
jmdn. auf den Arm nehmen	19	31	50
das Eis brechen	3	82	85
etw. auf Eis legen	1	49	50
die Fäden ziehen	9	189	198
aufs falsche Pferd setzen	2	55	57
mit dem Feuer spielen	8	86	94
gegen den Strom schwimmen	1	60	61
die Kastanien aus dem Feuer holen	0	46	46
in den Keller gehen	28	63	91
im Regen stehen	20	80	100
den richtigen Ton treffen	30	80	110
in Stein gemeißelt sein	8	4	12
unter den Teppich kehren	0	75	75
ins Wasser fallen	46	124	170
das Wasser bis zum Hals stehen haben	17	75	92
total	192	1099	1291

Table 1: Instances of MWEs pulled form the corpus.

The annotation process revealed a considerable limitation of the cohesion-based method that was also mentioned by Sporleder and Li (2009): If the idiomatic reading is not isolated, but is lexically

[7] PMI is an association measure of two words. It is the ratio of the probability that the two words occur together to the probability that the two words appear independent of each other.

[8] *Tübingen Partially Parsed Corpus of Written German*

[9] Nunberg et al. point out that although "speakers may not always perceive the precise motive for the figure involved [...] they generally perceive **that** some form of figuration is involved" (1994, p. 492).

cohesive with regard to its context, the method obviously has to fail. But when does this happen? There were a few cases where an idiom did not stick out, because a whole metaphorical context was created around it. For example, one instance of the MWE *aufs falsche Pferd setzen* ('to back the wrong horse') was used together with other terms of the domain equitation to depict an unfortunate politician as a rider who falls from his horse. And sometimes it was the other way round. Some authors deliberately played with the ambiguity of an MWE by using it in a literal context with an idiomatic meaning (for example the fish who *swam against the tide*). Unfortunately this is a limitation one cannot overcome when using a cohesion-based method.

For the identification task a classifier was implemented that was based on the cohesion graphs of Sporleder and Li. An example for a cohesion graph is shown in Figure 1. In these undirected graphs nodes correspond to words and each node is connected with all other nodes. The edges are labeled with the cosine of the corresponding vectors. The cosine of an angle between two vectors is indicative for the semantic similarity of the words representend by those vectors. The larger the cosine (i.e. the smaller the angle), the more similar are these terms.

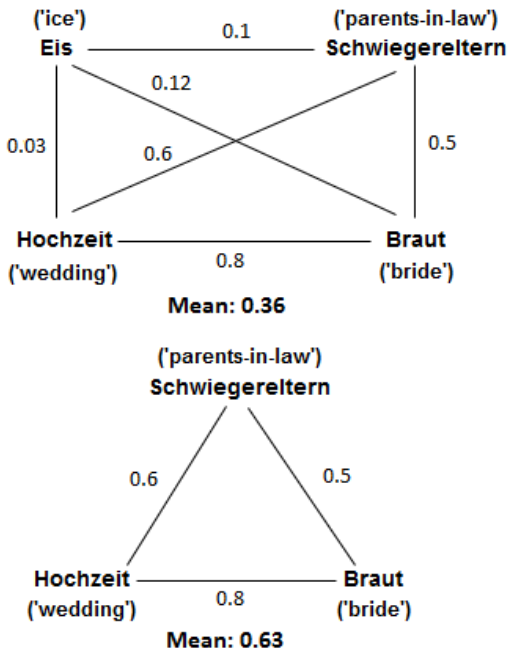

Figure 1: Example of two cohesion graphs with their respective mean cosine distance.

Figure 1 illustrates the identification process for example (2).[10] The graph at the top still contains the noun *Eis* component of the idiom *das Eis brechen*[11] and has connectivity mean of 0.36. In the graph at the bottom *Eis* was removed and the connectivity rose to a mean of 0.63. Since the cohesion between the words in the graph has increased, this is a sign for an idiomatic reading of the MWE.

The identification task was conducted as follows: First the paragraphs containing the instances of MWEs were reduced to only nouns (this will be explained later). Then the noun component of the MWE and a fixed number of neighbouring words were used to build a graph like in Figure 1. The similarity values were calculated by assigning the vector representations to the words from a vector lexicon and then calculating the cosine values of these vectors. After completing the graph the mean of the cosine values was calculated. After this the noun component of the MWE was removed from the graph and the mean was calculated again. If the mean got larger, the classifier labeled the instance of the MWE as *idiomatic*, if it stayed the same or got smaller, the instance was labeled as *literal*.

To test the impact of the different approaches on the representation of semantic similarity both types of VSMs, unstructured and structured, were employed in the experiments. Because the unstructured model did outperform the structured one, another unstructured model was built using different parameters to check, whether the performance could further be enhanced. Thus, a total of three different vector lexicons were used.

The first vector lexicon used was the German version of the Distributional Memory framework (DM) by Padó and Utt (2012). DM, originally designed for English by Baroni and Lenci (2010), is a structured distributional semantics model that includes grammatical dependencies. In contrast to the common approach to collect the data in a matrix, DM gathers it in a third-order tensor[12], i.e. in form of weighted word-link-word tuples (for ex-

[10] The nodes correspond to the nouns in (2). Since the experiments were conducted on the basis of a German corpus, the node labels are the respective German terms for *ice*, *parents-in-law*, *wedding* and *bride*.

[11] 'to break the ice'

[12] Tensors are generalisations of vectors and matrices. A first-order tensor is a vector, a second-order tensor a matrix and a third order tensor a three-dimensional array (Erk, 2012).

ample (*soldier*, *sbj_intr*, *talk* 5.42)). The tensor makes it possible to create different matrices on demand: word × link-word, word-word × link, word-link × word and link x word-word. For the purpose of this experiment a word × link-word matrix was generated since we want to compare the semantic similarity of single words. Then singular value decomposition (SVD)[13] was applied to the matrix to reduce the dimensions of the word vectors to 300.

The second vector lexicon was created with the word2vec tool on the basis of DECOW14, a German gigatoken web corpus provided by the COW (COrpora from the Web) initiative led by Felix Bildhauer and Roland Schäfer at Freie Universität Berlin (Schäfer and Bildhauer, 2012). The word embeddings generated by word2vec had a dimensionality of 100.

Last but not least, a third vector lexicon was created using the hyperwords tool provided by Omer Levy, also with the DECOW14 corpus as a basis. This tool incorporates the lessons learned of Levy et al. (2015) which were shortly presented in section 3.2. The word embeddings generated by hyperwords had 500 dimensions.

The decision to only include nouns in the identification process was made to significantly reduce the size of the vector lexicons and thereby the computational costs. Nouns were chosen, because they are considered to be the best topic indicators in a text.

All three vector lexicons were tested with a window of size six around the MWE.[14] Subsequently the best performing vector lexicon was tested with context windows of size two and size ten to examine the effect of the window size on the performance.

4 Results

The baseline for the experiments was a classifier that labeled all instances with the majority class. Thus, the accuracy, for example, would be 85.13% because 85.13% of the instances are idiomatic. Since we made the assumption that word embeddings are better suited for the presented method than the NGD, the NGD would of course have been a more natural baseline. Unfortunately, getting the required data proved to be not that easy because the access to the search APIs of the major search engines seems to be more restricted than a few years ago.

Table 4 shows the results for the three vector lexicons with a context window of 6. With an accuracy of 63.35% DM showed by far the worst performance, falling short of the baseline by a large margin. The reason might be that while the NGD only considers syntagmatic relations between words (i.e. the question if they co-occur), DM seems to have its focus on paradigmatic relations. This would explain why words like *France - Italy* (0.84)[15], *president - Pope* (0.78) and *minister of defence - general* (0.77) are pretty close in this word space, whereas terms like *murder - court* (0.058), *president - USA* (0.078) and *city - border* (0.047) are very far apart, though clearly topically related. Words that build a paradigm exhibit a substitutional relationship which means that one word can potentially replace the other in a specific context (e.g. *The president/Pope gave a speech.*). And if a word can be replaced by another this in turn means that they have to be attributionally similar which appears to be exactly the kind of similarity DM represents. This is bad news for the task at hand, since lexical cohesion, as we saw, not only incorporates attributional similarity, but all kinds of relations. However, words that are connected by a nonsystematic relationship are very dissimilar to each other according to DM. This could indicate that structural distributional semantics models (at least the ones that rely on grammar dependencies) are not the best solution for cohesion-based tasks.

Word2vec on the other hand delivered with an accuracy of 81.03% the best performance for a context window size of 6 (but still falling below the baseline by ca. 4%). This is in accordance with the above presented suspicion that a structured model is not a good fit for the conducted experiments. After all word2vec respectively skip-gram (which was used for the experiment) is an unstructered model. In contrast to DM, word2vec not only seems to model attributional similarity, e. g. *Apfel - Birne*[16] (0.8), but also topical relat-

[13] SVD is a dimensionality reduction technique. Through SVD a matrix is decomposed in three matrices whose dimensions are reduced to a desired number. The matrices originating from this process approximate the original matrix. This is possible because the remaining dimensions are the principal components of the data, i.e. they convey the most information.

[14] The number of neighbouring words that were included in the cohesion graphs along with the noun component of the idiom.

[15] In the parentheses behind the word pairs are the cosine values.

[16] *Apfel* means 'apple', *Birne* means 'pear'.

	DM			Word2vec			Hyperwords		
MWE	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**
jmdn. auf den Arm nehmen	53.33	26.67	39.58	81,25	86,67	79,20	61,34	90,00	58,33
das Eis brechen	95.56	55.13	54.32	98,53	85,60	85,20	97,33	93,59	91,36
etw. auf Eis legen	100	50.00	51.06	97,87	100	98,00	97,87	100	97,87
die Fäden ziehen	96.93	86.81	84.82	97,42	82,51	81,25	96,49	90,66	87,96
aufs falsche Pferd setzen	93.94	62.00	59.62	98,08	100	98,10	96,23	100	96,23
mit dem Feuer spielen	91.25	91.25	84.09	95,89	87,50	85,23	97,06	82,50	81,82
gegen den Strom schwimmen	98.15	92.98	91.38	100	87,93	88,14	100	69,64	70,18
die Kastanien aus dem Feuer holen	100	100	100	100	68,89	68,89	100	68,89	68,89
in den Keller gehen	71.43	8.06	32.18	85,94	88,71	81,61	78,67	95,16	78,16
im Regen stehen	80.60	72.00	64.21	87,80	93,51	84,54	87,01	88,16	80,21
den richtigen Ton treffen	77.78	18.42	37.14	82,43	78,21	71,96	82,09	71,43	67,92
in Stein gemeißelt sein	50.00	25.00	63.64	42,86	75,00	58,33	37,5	75,00	50,00
unter den Teppich kehren	100	45.59	45.59	100	92,86	92,86	100	91,30	91,30
ins Wasser fallen	74.10	86.55	67.90	85,22	81,67	76,22	81,89	87,39	76,69
das Wasser bis zum Hals stehen haben	81.82	88.73	74.71	92,00	63,89	65,91	86,30	87,50	78,41
total	84.33	60.61	63.35	89,69	84,86	81,03	86.65	86,08	78,36

Table 2: Results for the three different vector spaces with a context window of size 6.

edness as is shown in Figure 1. A *wedding* and a *bride* do not have much in common in terms of their properties (one is an event, the other is a human being), but they are undoubtedly topically close as word2vec correctly assumes (0.8).

The performance of hyperwords (78.63% accuracy) is comparable to that of word2vec which is not very surprising since it also uses SGNS only with different parameter settings.[17]

The best model, word2vec, was then used to examine the effect of different context window sizes on the performance. At first, a very narrow window of size 2 was tested to check whether the two closest neighbours[18] are sufficient to identify the idiomatic reading of an MWE. The results seen in table 3 suggest they are not. With 63.26% accuracy it performs as badly as the DM model with a context window of 6.

Subsequently a broader window of size 10 was used while conducting the task. In contrast to the narrow window it performed well and achieved with an accuracy of 85.67% (see table 4) the highest score of the experiment, surpassing the accuracy baseline by a slight bit. But since we want our classifier to perform well on both classes, idiomatic and literal, it is important to also have a look at the precision (90.47%) which surpasses the baseline by more than 5%. The good performance

[17]Hyperwords offers two different possibilites: the 'old way' of creating a word-context matrix reduced with SVD, and SGNS. We used SGNS in the experiments.

[18]Reminder: The noun component of the MWE is in the focus of the window.

MWE	**Pre**	**Rec**	**Acc**
jmdn. auf den Arm nehmen	76,00	61,29	64,00
das Eis brechen	98,25	70,00	69,88
etw. auf Eis legen	97,78	89,80	88,00
die Fäden ziehen	96,50	58,20	58,08
aufs falsche Pferd setzen	95,56	78,18	75,44
mit dem Feuer spielen	98,11	61,90	64,13
gegen den Strom schwimmen	100	63,33	63,93
die Kastanien aus dem Feuer holen	100	8,89	8,89
in den Keller gehen	85,71	76,19	74,73
im Regen stehen	88,57	77,50	74,00
den richtigen Ton treffen	85,48	66,25	67,27
in Stein gemeißelt sein	60,00	75,00	75,00
unter den Teppich kehren	100	72,97	72,97
ins Wasser fallen	84,54	66,13	66,47
das Wasser bis zum Hals stehen haben	81,82	12,00	26,09
total	89,89	62,51	63,26

Table 3: Results for the word2vec vector space with a context window of size 2.

compared to the other results indicates a correlation between the size of the context window and the performance of the model.

5 Conclusion

The experiments conducted in the course of this work show that the presented method generally produces good results if a suitable vector lexicon is used and the context window is large enough. These results could probably further be improved if different parameters are optimized. It is possible that the model would achieve even better results by including verbs in the cohesion graphs in addition to nouns since they are also good topic indicators. In addition, it would be interesting to see

MWE	Pre	Rec	Acc
jmdn. auf den Arm nehmen	86,67	92,86	86,36
das Eis brechen	100	88,89	89,33
etw. auf Eis legen	97,73	100	97,73
die Fäden ziehen	97,30	86,75	85,14
aufs falsche Pferd setzen	97,83	100	97,83
mit dem Feuer spielen	95,65	90,41	87,65
gegen den Strom schwimmen	97,67	91,30	89,36
die Kastanien aus dem Feuer holen	100	85,37	85,37
in den Keller gehen	87,30	94,83	86,42
im Regen stehen	86,84	97,06	85,71
den richtigen Ton treffen	85,71	89,55	81,52
in Stein gemeißelt sein	50,00	75,00	60,00
unter den Teppich kehren	100	96,77	96,77
ins Wasser fallen	82,57	83,33	74,66
das Wasser bis zum Hals stehen haben	91,80	84,85	81,25
total	90,47	90,46	85,67

Table 4: Results for the word2vec vector space with a context window of size 10.

up to which point an enlargement of the context window results in a better performance.

For further future work, it would be desirable to test if the method could be used to automatically discover non-compositional MWEs when combined with a statistical approach. First, with help of a measure of association one could generate a candidate list of statistically idiomatic MWEs whose instances are then examined for lexical cohesion with respect to their contexts. This way, it may be possible to discriminate between institutionalized phrases and non-compositional MWEs.

6 Acknowledgements

I am grateful to my supervisors Laura Kallmeyer and Timm Lichte for their valuable feedback and guidance throughout this work. In addition I would like to thank Sebastian Padó, Jason Utt, Felix Bildhauer and Roland Schäfer for the provided ressources and the three anonymous reviewers for their comments on this paper.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

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

John Rupert Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, pages 491–538.

Sebastian Padó and Jason Utt. 2012. A distributional memory for german. In Jeremy Jancsary, editor, *KONVENS*, volume 5, pages 462–470. ÖGAI, Wien, Österreich.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul. ELRA.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece, March. Association for Computational Linguistics.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton. 2004. Select: a lexical cohesion based news story segmentation system. *AI Communications*, 17(1):3–12.