A Single Word is not Enough: Ranking Multiword Expressions Using Distributional Semantics

Martin Riedl and **Chris Biemann**
Language Technology
Computer Science Department, Technische Universität Darmstadt
Hochschulstrasse 10, D-64289 Darmstadt, Germany
{riedl, biem}@cs.tu-darmstadt.de

Abstract

We present a new unsupervised mechanism, which ranks word n-grams according to their multiwordness. It heavily relies on a new *uniqueness* measure that computes, based on a distributional thesaurus, how often an n-gram could be replaced in context by a single-worded term. In addition with a downweighting mechanism for incomplete terms this forms a new measure called DRUID. Results show large improvements on two small test sets over competitive baselines. We demonstrate the scalability of the method to large corpora, and the independence of the measure of shallow syntactic filtering.

1 Introduction

While it seems intuitive to treat certain sequences of tokens as single terms, there is still considerable controversy about the definition of what exactly such a multiword expression (MWE) constitutes. Sag et al. (2001) pinpoint the need of treating MWEs correctly and classify a range of syntactic formations that could form MWEs and define MWEs as being non-compositional with respect to the meaning of their parts. While the exact requirements on MWEs is bound to specific tasks (such as parsing, keyword extraction, etc.), we operationalize the notion of non-compositionality by using distributional semantics and introduce a new measure that works well for a range of task-based MWE definitions.

Most previous MWE ranking approaches use the following mechanisms to determine multiwordness: part-of-speech (POS) tags, word/multiword frequency and significance of co-occurrence of the parts. In this paper we do not want to introduce ”yet another ranking function” but rather an additional mechanism, which performs ranking based on distributional semantics.

Distributional semantics has already been used for MWE identification, but mainly to discriminate between compositional and non-compositional MWEs (Schone and Jurafsky, 2001; Salehi et al., 2014; Hermann and Blunsom, 2014). Here we introduce a new concept to describe the multiwordness of a term by its *uniqueness*. Using the *uniqueness* score we measure how likely a term in context can be replaced by a single word. This measure is motivated by the semiotic consideration that due to parsimony concepts are often expressed as single words. Furthermore, we implement a context-aware punishment, called *incompleteness*, which degrades the score of terms that seem incomplete regarding their contexts. Both concepts are combined into a single score we call DRUID, which is calculated based on a distributional thesaurus. In the following, we show the impact of that new method for French and English and also examine the effect of corpus size on MWE extraction. Additionally, we report on results without using any linguistic preprocessing except tokenization.

2 Related Work

The generation of MWE dictionaries has drawn much attention in the field of Natural Language Processing (NLP). Early computational approaches (e.g. Justeson and Katz (1995)) use POS sequences as MWE extractors. Other approaches, relying on word frequency, statistically verify the hypothesis whether the parts of the MWE occur more often together than would be expected by chance (Manning and Schütze, 1999; Evert, 2005; Ramisch, 2012). One of the first measures that consider context information (co-occurrences) are the C-value and the NC-value introduced by Frantzi et al. (1998). These methods first extract candidates using POS information and then compute scores based on the frequency of the MWE and the frequency of nested MWE candidates. The method described by Wermter and Hahn (2005) computes a score by multiplying the frequency of a candidate when placing wildcards for each word. A newer method is introduced in Lossio-Ventura et al. (2014), which reranks scores based on an extension of the C-value, which uses a POS-based probability and an inverse document frequency. Using different measures and learning a classifier that predicts the multiwordness was first proposed by Pecina (2010), who, however, restricts his experiments to two-word MWEs for the Czech language only. Korkontzelos (2010) comparatively evaluates several MWE ranking measures. The best MWE extractor reported in his work is the scorer by (Nakagawa and Mori, 2002; Nakagawa and Mori, 2003), who use the un-nested frequency (called marginal frequency) of each candidate and multiply these by the geometric mean of the distinct neighbor of each word within the candidate.

Distributional semantics is mostly used to detect compositionality of MWEs (Salehi et al., 2014; Katz and Giesbrecht, 2006). Most approaches therefore compare the context vector of a MWE with the combined vectors based on the constituent words of the MWE. The similarity between the vectors is then used as degree for compositionality. In machine translation, words are sometimes considered as multiwords if they can be translated as single term (cf. (Bouamor et al., 2012; Anastasiou, 2010)). Whereas this follows the same intuition as our *uniqueness* measure, we do not require any bilingual corpora.

Regarding the evaluation, mostly precision at k ($P@k$) and recall at k ($R@k$) are applied (e.g. (Evert, 2005; Frantzi et al., 1998; Lossio-Ventura et al., 2014)). Another general approach is using the average precision (AP), which is also used in Information Retrieval (IR) (Thater et al., 2009) and has also been applied by Ramisch et al. (2012).

3 Baselines and Previous Approaches

We will evaluate our method by comparing our MWE ranking to multiword lists that have been annotated in corpora. Here, we introduce an upper bound and two baseline methods and give a brief description of the competitive methods used in this paper. Most of these methods require a list of candidate terms T, usually extracted with POS sequences (see Section 5).

3.1 Upper Bound

We use a perfect ranking as upper bound, where we rank all positive candidates before all negative ones.

3.2 Lower Baseline and Frequency Baseline

The ratio between true candidates and all candidates serves as lower baseline, which is also called baseline precision (Evert, 2008). The second baseline is the frequency baseline, which ranks candidate terms $t \in T$ according to their frequency $freq(t)$.

3.3 C-value/NC-value

The commonly used C-value (see Eq. 1) was developed by Frantzi et al. (1998). The first factor, logarithm of the term length in words, favors longer MWEs. The second factor is the frequency of the term reduced by the average frequency of all candidate terms T, which nest the term t, i.e. t is a substring of the terms we denote as T_t.

$$\mathrm{c}(t) = \log_2(|t|)(\mathrm{freq}(t) - \frac{1}{|T_t|}\sum_{b \in T_t} \mathrm{freq}(b)) \quad (1)$$

An extension of the C-value was proposed by Frantzi et al. (1998) as well and is named NC-value. It takes advantage of context words C_t by assigning weights to them. As context words only *nouns*, *adjectives* and *verbs* are considered[1]. Context words are weighted with Equation 2, where k denotes the number of times the context word $c \in C_t$ occurs with any of the candidate terms. This number is normalized by the number of candidate terms.

$$w(c) = \frac{k}{|T|} \quad (2)$$

The NC-value is a weighted sum of the C-value and the product of the term t occurring with each context c which form the term t_c:

$$\mathrm{nc}(t) = 0.8 * \mathrm{c}(t) + 0.2 \sum_{c \in C_t} freq(t_c) w(c). \quad (3)$$

3.4 t-test

The t-test (see e.g. (Manning and Schütze, 1999, p.163)) is a statistical test for the significance of

[1] the context window size is not specified in Frantzi et al. (1998)

co-occurrence of two words. It relies on the probabilities of the term and its single words. The probability of a word $p(w)$ is defined as the frequency of the term divided by the total number of terms of the same length. The *t-test* statistic is computed using Equation 4 with $freq(.)$ being the total frequency of unigrams.

$$t(w_1 \dots w_n) \approx \frac{p(w_1 \dots w_n) - \prod_{i=1}^{n} p(w_i)}{\sqrt{p(w_1 \dots w_n)/freq(.)}} \quad (4)$$

We then use this score to rank the candidate terms.

3.5 FGM Score

Another method inspired by the C/NC-value is proposed in (Nakagawa and Mori, 2002; Nakagawa and Mori, 2003). The method was developed on a Japanese dataset and outperformed a modified C-value[2] measure. The method is composed of two scoring mechanisms for the candidate term t as shown in Equation 5.

$$FGM(t) = GM(t) \times MF(t) \quad (5)$$

The first term in the equation is a geometric mean $GM(.)$ of the number of distinct direct left $l(.)$ and right $r(.)$ neighboring words for each single word t_i within t.

$$GM(t) = \sqrt[2|t|]{\sum_{t_i \in t} (|l(t_i)| + 1)(|r(t_i)| + 1)} \quad (6)$$

The neighboring words are extracted directly from the corpus; the method does neither rely on candidate lists nor POS tags. To the contrary, the marginal frequency $MF(t)$ relies on the candidate list and the underlying corpus. This frequency counts how often the candidate term occurs within the corpus and is not a subset of a candidate. In Korkontzelos (2010) it was shown that while scoring according to Equation 5 leads to comparatively good results, it is consistently outperformed by MF only.

4 Semantic Uniqueness and Incompleteness

We present two new mechanisms relying on a Distributional Thesaurus (DT), which we use to rank terms regarding their multiwordness: A score for the *uniqueness* of a term and a punishing score that conveys the *incompleteness*.

4.1 Similarity Computation

The DT is computed based on Biemann and Riedl (2013). First we extract n-grams from text and consider the left and the right neighbor of each n-gram as context feature. Then, we calculate the Lexicographer's Mutual Information (LMI) significance score (Bordag, 2008) between n-grams and features and remove all context features, which co-occur with more than 1000 terms, as these features tend to be to general. In the next step we keep for each n-gram only the 1000 context features, with the highest LMI score. The similarity score is then computed based on the overlap of features between two terms. Due to pruning this overlap-based significance measure is proportional to the Jaccard similarity measure, albeit we do not consider any normalization. After computing the feature overlap between two terms, we keep for each n-gram the 200 most similar n-grams. An example for the most similar n-grams to the terms *red blood cell* and *red blood* including their feature overlap are shown in Table 1.

4.2 Uniqness Computation

The first mechanism of our MWE ranking method is based on the following hypothesis: n-grams, which are MWE, could be substituted by single words, thus they have many single words amongst their most similar terms. This is motivated by semiotic considerations: Because of parsimony, concepts are usually expressed in single words. When a semantically non-compositional word combination is added to the vocabulary, it expresses a concept that is necessarily similar to other concepts. Hence, if a candidate multiword is similar to many single word terms, this indicates multiwordness.

To compute the *uniqueness* score (uq) of an n-gram t, we first extract the n-grams it is similar to using the DT as described in Section 4.1. The function $similarities(t)$ returns the 200 most similar n-grams to the given n-gram t. We then compute the ratio between unigrams and all similar n-grams considered using the formula:

$$uq(t) = \frac{\sum_{s:|s|=1}^{similarities(t)} 1}{|similarities(t)|}. \quad (7)$$

We illustrate the computation of our measure based on the MWE *red blood cell* and the non-MWE *red blood*. When considering only the ten most similar entries for both n-grams as illustrated

[2] They adjust the logarithmic length in order to be able to use the C-value to detect single worded terms.

in Figure 1, we observe an uniqueness score of $7/10 = 0.7$ for both n-grams. If considering the top 200 similar n-grams, which are also used in our experiments we will obtain 135 unigrams for the candidate *red blood cell* and 100 unigrams for the n-gram *red blood*. We will use these counts for showing the workings of the method in the remainder.

red blood cell		*red blood*	
Sim. term	Sc.	Sim. term	Sc.
erythrocyte	133	**red**	148
red cell	129	white blood	111
RBC	95	**Sertoli**	93
platelet	70	**Leydig**	92
red-cell	37	**NK**	86
reticulocyte	34	**mast**	85
white blood	33	**granulosa**	81
leukocyte	29	**endothelial**	81
granulocyte	28	hematopoietic stem	79
the erythrocyte	28	peripheral blood monon	78

Table 1: We show the ten most similar entries for the term *red blood cell* (left) and *red blood* (right). Here, seven out of ten terms are single words.

4.3 Incompleteness Computation

Similar to the C/NC-value method, we also assign a context weighting function that punishes incomplete terms, which we call *incompleteness (ic)*. For this function we extract the 1000 most significant context features using the function $context(t)$, which yields tuples of left and right contexts. These context features are the same that are used for the similarity computation in Section 4.1 and have been ranked according to the LMI measure. For the example term *red blood*, some of the contexts are $\langle$*extravasated, cells*$\rangle$, $\langle$*uninfected, cells*$\rangle$, $\langle$*nucleated, corpuscles*$\rangle$. In the next step we split each tuple to its left and right word including its relative position (left/right) to the candidate term. Using the first context feature results in: $\langle$*extravasated, left*$\rangle$, $\langle$*cells, right*$\rangle$. Then, we sum up the occurrences of for each single context, as shown in Table 2 for the two terms.

We subsequently select the maximal count and normalize it by the counts of features $|context(t)|$ considered, which is 1000. This results into the incompleteness measure $ic(t)$. For our example terms we achieve the values $ic(red\ blood) = 557/1000$ and $ic(red\ blood\ cell) = 48/1000$. Whereas the uniqueness scores for the most similar entries were equal, we now have a measure that indicates the incompleteness of an n-gram, with higher scores indicating more incomplete terms.

Context term	Position	Count
red blood cell		
transfusions	right	48
(	right	42
transfusion	right	33
red blood		
cells	right	557
cell	right	344
corpuscles	right	13

Table 2: Top three most frequent context words for the term *red blood cell* and *red blood* in the Medline corpus.

4.4 Combining Both Measures

As shown in the previous two sections, a high uniqueness score indicates the multiwordness and a high incompleteness score should decrease the overall score. In experiments, we found the best combination if we subtract[3] the incompleteness score from the uniqueness score. This mechanism is inspired by the C-value and motivated as terms that are often preceded/followed by the same word do not cover the full multiword expression and need to be downranked. This leads to Equation 8, which we call **D**ist**R**ibutional **U**niqueness and **I**ncompleteness **D**egree:

$$\texttt{DRUID}(t) = \mathrm{uq}(t) - \mathrm{ic}(t). \qquad (8)$$

Applying the DRUID score to our example terms (considering the 200 most similar terms) we will achieve the scores $\texttt{DRUID}(red\ blood\ cell) = 135/200 - 48/1000 = 0.627$ and $\texttt{DRUID}(red\ blood) = 100/200 - 557/1000 = -0.057$. As a higher DRUID score indicates the multiwordness of an n-gram, we can summarize that the n-gram *red blood cell* is a better MWE than the n-gram *red blood*.

5 Experimental Setting

We examine two experimental settings: First, we compute all measures on a small corpus that has been annotated for MWEs, which serves as the gold standard. In the second setting we compute the measures on a larger in-domain corpus. The evaluation is again performed for the same candidate terms as given by the gold standard. Results for the top k ranked entries are reported using the precision at k ($P@k = \frac{1}{k}\sum_{i=1}^{k} x_i$ with x_i equals 1 if the i-th ranked candidate is annotated as MWE and 0 otherwise). For an overall performance we

[3] multiplicative combinations consistently performed worse

use the average precision (AP) as defined in Thater et al. (2009): $AP = \frac{1}{|T_{mwe}|} \sum_{k=1}^{|T|} x_k P@k$, with T_{mwe} beeing the set of positive MWE. When facing tied scores we mix false and true candidates randomly cf. Cabanac et al. (2010).

5.1 Corpora

For the experiments we consider two annotated (small) corpora and two unannotated (large) corpora.

5.1.1 GENIA corpus and SPMRL 2013: French Treebank

In the first experiments we use two small annotated corpora that serve the gold standard MWEs. We use the medical GENIA corpus (Kim et al., 2003)[4] which consists of 1999 abstracts from Medline[5] and encompasses 0.4 million words. This corpus has annotations regarding important and biomedical terms. Also single terms are annotated in this data set, which we ignore.

The second small corpus is based on the French Treebank (Abeillé and Barrier, 2004), which was extended for the SPMRL task (Seddah et al., 2013). This version of the corpus also contains compounds annotated as MWEs. In our experiments we use the training data, which covers 0.4 million words.

Whereas the GENIA MWEs target term matching and medical information retrieval, the SPMRL MWEs mainly focus on improving parsing through compound recognition.

5.1.2 Medline Corpus and Est Républican Corpus (ERC)

In a second experiment the scalability to larger corpora is tested. For this, we make use of the entire Medline[5] abstracts, which consist of about 1.1 billion words. The Est Républican Corpus (ERC) (Seddah et al., 2012)[6] is our large French corpus. It consists of local French news from the eastern part of France and comprises of 150 million words.

5.2 Candidate Selection

In the first two experiments, we use POS filters to select candidates. We concentrate on filters that extract noun MWEs and avoid further preprocessing like lemmatization. We use the filter introduced by Justeson and Katz (1995)[7] for the English medical datasets. Considering only terms that appear more than ten times leads to 1,340 candidates for the GENIA dataset and 29,790 candidates for the Medline dataset. According to Table 3 we observe that most candidates are bigrams. Whereas for both corpora still around 20% of trigrams are contained, the number of 4-grams is only marginally represented. For the French datasets we apply the filter proposed by Daille et al. (1994)[8], which is suited to match nominal MWEs. Applying the same filtering as for the medical corpora leads to 330 candidate terms for the SPMRL and 7,365 candidate terms for the ERC. Here the ratio between bi- and trigrams is more balanced but again the number of 4-grams constitutes the smallest class.

Corpus	Candidates	2-gram	3-gram	4-gram
GENIA	1,340	1,056	243	41
Medline	29,790	22,236	6,400	1,154
SPMRL	330	197	116	17
ERC	7,365	3,639	2,889	837

Table 3: Number of candidates after filtering for the expected POS-tag and their distribution over n-grams with $n \in \{1, 2, 3, 4\}$.

In comparison to the Medline dataset, the ratio of multiwords extracted by the POS filter on the French corpus is much lower. The reason for that property is that in the French data, many adverbial, prepositional MWEs are annotated, which are not covered by the POS filter.

The third experiment shows the performance of the method in absence of language-specific preprocessing. Thus, we only filter the candidates by frequency and do not make use of POS filtering. As most previous methods rely on POS-filtered data we cannot make use of them in this setting.

For the evaluation, we compute the scores of the competitive methods in two settings: First, we compute the scores based on the full candidate list without any frequency filter and prune low-frequent candidates only for the evaluation (post-prune). In the second setting we filter candidates

[4] freely available at http://www.nactem.ac.uk/genia/genia-corpus/pos-annotation.

[5] http://www.nlm.nih.gov/bsd/licensee/access/medline_pubmed.html

[6] http://www.cnrtl.fr/corpus/estrepublicain

[7] A regular expression for matching POS tag sequences is summarized by Korkontzelos (2010): `(([JN]+[JN]?[NP]?[JN]?)N)`. Each letter is a truncated POS tag of length one where J is an adjective N a noun and P a preposition.

[8] Following the same convention as for English the regular expression can be expressed as `N[J]?|NN|NPDN`

according to their frequency before the computation of scores (pre-prune). This leads to differences for context-aware measures, since in the pre-pruned case, a lower number of less noisier contexts is used.

6 Results

6.1 Small Corpora Results

First we show the results based on the GENIA corpus (see Table 4). Almost all competitive methods beat the lower baseline. The C/NC-value perform best when the pruning is done after a frequency filter. In line with the findings of Korkontzelos (2010) and in contrast to Frantzi et al. (1998) the AP of the C-value is slightly higher than for the NC-value. All the FGM based methods except the GM measure alone outperform the C-value. The results in Table 4 indicate that the best competitive system is the post-pruned FGM system as it has much higher average precision scores and misses only 50 MWEs in the first 500 entries. A slightly different picture is presented in Figure 1 where the $P@k$ scores against the number of candidates are plotted. Here DRUID performs well for the top-k list for small k, i.e. finds many good MWEs with high confidence thus combines well with MF, which extends to larger k, but places too much importance of frequency when used alone. Common errors are frequent chunks such as ”in patience”, see Table 9 in Section 7. Whereas for the post-pruned case FGM scores higher than MF, the inverse is observed when pre-pruning. Using our DRUID methods can surmount the FGM method only for the first 300 ranked terms (see Figure 1 and Table 4). Multiplying the logarithmic frequency to the DRUID, the results improve slightly and the best $P@100$ with 0.97 is achieved. All FGM results are outperformed when combining the post-pruned FGM scores with our measure. According to Figure 1 this combination achieves high precision for the first ranked candidates and still exploits the good performance of the post-pruned FGM based method for the middle-ranked candidates.

Method	$P@100$	$P@500$	AP
upper baseline	1.000	1.000	1.0000
lower baseline	0.713	0.713	0.7134
frequency	0.790	0.750	0.7468
t-test	0.790	0.750	0.7573
C-value (pre-pruned)	0.880	0.846	0.8447
NC-value (pre-pruned)	0.880	0.840	0.8405
GM	0.590	0.662	0.6740
MF (pre-pruned)	0.920	0.872	0.8761
FGM (pre-pruned)	0.910	0.840	0.8545
MF (post-pruned)	0.900	0.876	0.8866
FGM (post-pruned)	0.900	0.900	0.8948
DRUID	0.930	0.852	0.8663
log(freq)(DRUID)	**0.970**	0.860	0.8661
MF(post-pruned)DRUID	0.950	0.926	**0.9241**
FGM(post-pruned)DRUID	**0.960**	**0.940**	**0.9262**

Table 4: Results for the GENIA dataset.

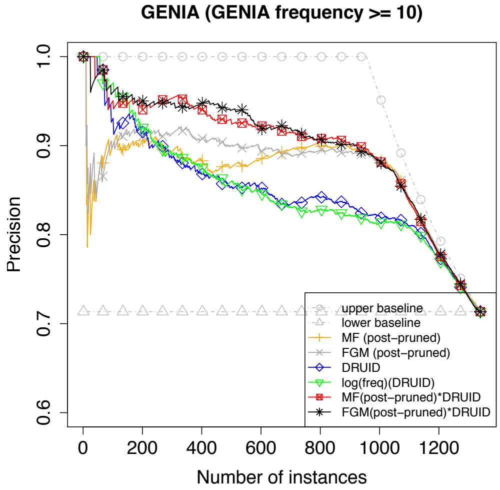

Figure 1: Results for the GENIA corpus.

Different results are achieved for the SPMRL dataset as can be seen in Table 5. Whereas the pre-pruned C-value again receives better results than frequency, it scores below the lower baseline. Also the post-pruned FGM and MF method do not exceed the lower baseline. Data analysis revealed that for the French dataset only ten out of the 330 candidate terms are nested within any of the candidates. This is much lower than the 637 terms nested in the 1340 candidate terms for the GENIA dataset. As both the FGM-based methods and the C/NC-value heavily rely on nested candidates, they cannot profit from the candidates of this dataset and achieve similar scores as ordering candidates according to frequency. Comparing the

Scoring	$P@100$	$P@200$	AP
upper baseline	1.000	0.860	1.0000
lower baseline	0.521	0.521	0.5212
frequency	0.500	0.480	0.4876
t-test	0.500	0.485	0.4934
C-value (pre-pruned)	0.490	0.540	0.5107
MF (post-pruned)	0.510	0.495	0.5017
FGM (post-pruned)	0.460	0.480	0.4703
DRUID	**0.790**	**0.690**	**0.7794**
log(freq)DRUID	0.770	0.675	0.7631
MF(post-pruned)DRUID	0.700	0.630	0.6850
FGM(post-pruned)DRUID	0.600	0.570	0.5948

Table 5: Results for the French SPMRL dataset

baselines to our scoring method, this time we obtain the best result for DRUID without additional factors. However, multiplying DRUID with MF or log(frequency) still outperforms the other methods and the baselines.

6.2 Large Corpora Results

Most MWE evaluations have been performed on rather small corpora. Here we want to inspect the performance of the measures for large corpora, so as to realistically simulate a situation where the MWEs should be found automatically for an entire domain.

Using the Medline corpus, all methods except the GM score outperform the lower baseline and the frequency baseline (see Table 6). Regarding the AP the best results are obtained when combining our DRUID method with the MF, whereas for P@100 and P@500 the log-frequency weighted DRUID scores best. Using solely the DRUID method or the combined variation with the log-frequency lead to the best ranking for the first 1000 ranked candidates and is then outperformed by the MF based DRUID variations. In this experiment the C-value achieves the best performance from the competitive methods for the P@100 and P@500, followed by the t-test. But the highest AP is reached with the post-pruned MF method, which also outperforms the sole DRUID slightly. Contrary to the GENIA results, the MF scores are consistently higher than the FGM scores.

Scoring	P@100	P@500	AP
upper baseline	1.000	1.000	1.0000
lower baseline	0.416	0.416	0.4161
frequency	0.720	0.534	0.4331
C-value (pre-pruned)	0.750	0.564	0.4519
t-test	0.720	0.542	0.4483
GM	0.210	0.272	0.3502
MF (pre-pruned)	0.550	0.542	0.4578
FGM (pre-pruned)	0.580	0.478	0.4200
MF (post-pruned)	0.530	0.500	0.4676
FGM (post-pruned)	0.490	0.446	0.4336
DRUID	0.770	0.686	0.4608
log(freq)*DRUID	**0.860**	**0.720**	0.4693
GM*DRUID	0.770	0.634	0.4497
MF(pre-pruned)*DRUID	0.730	0.634	**0.4824**
MF(post-pruned)*DRUID	0.730	0.626	**0.4889**

Table 6: Results computed on the Medline corpus.

When using the French ERC we figured out that no nested terms are found within the candidates. Thus, the post- and pre-pruned settings are equivalent and thus MF equals frequency. The best results are again obtained with our method with and without the logarithmic frequency weighting (see Table 7). Again the AP of the C-value and most of the FGM-based methods are inferior to the frequency scoring. Only the t-test and the MF are slightly higher than the frequency[9]. But in contrast to the results based on the smaller SPMRL dataset, the MF, FGM and C-value can outperform the lower baseline. In comparison to the smaller corpora, the performance for the larger corpora is much lower. Especially low-frequent terms in the small corpora that have high frequencies in the larger corpora have not been annotated as MWEs.

Method	P@100	P@500	AP
upper baseline	1.000	1.000	1.0000
lower baseline	0.220	0.220	0.2201
frequency	0.370	0.354	0.3105
C-value	0.420	0.366	0.3059
t-test	0.390	0.360	0.3134
GM	0.010	0.052	0.1694
MF	0.370	0.356	0.3148
FGM	0.280	0.260	0.2405
DRUID	0.700	0.568	0.3962
log(freq)DRUID	**0.760**	**0.582**	**0.4075**
MF*DRUID	0.570	0.516	0.3776
FGM*DRUID	0.510	0.418	0.3234

Table 7: Results computed based on the ERC.

6.3 Results without POS Filtering

In the last experiment, we apply our method to candidates without any POS filtering and report results using a frequency threshold of ten. As the competitive methods from the previous section rely on POS tags, we use the t-test for comparison. Analysis revealed that the top-scored candidates according to the t-test begin with stop words.

		Medical		French	
	Method	P@100	AP	P@100	AP
small corpora	upper baseline	1.000	1.0000	1.000	1.0000
	lower baseline	0.107	0.1071	0.083	0.0832
	frequency	0.150	0.1135	0.060	0.0906
	t-test	0.160	0.1261	0.080	0.1097
	t-test + sw	0.530	0.3643	0.180	0.1481
	DRUID	**0.700**	**0.4048**	**0.670**	**0.2986**
	log(freq)DRUID	0.690	0.3644	0.460	0.2527
large corpora	upper baseline	1.000	1.0000	1.000	1.0000
	lower baseline	0.036	0.0361	0.019	0.0191
	frequency	0.010	0.0361	0.060	0.0366
	t-test	0.020	0.0412	0.080	0.0440
	t-test + sw	0.000	0.0989	0.200	0.0485
	DRUID	0.610	0.1378	**0.660**	**0.1009**
	log(freq)DRUID	**0.760**	**0.1649**	0.600	0.0988

Table 8: Results without linguistic pre-processing.

[9] This is achieved by chance for the MF, as it is equal to the frequency. The different scores are due to the randomly sorted tied scores used during our evaluation and reflect the variance of the randomness.

As an additional heuristic for the t-test, we shift MWEs, which start or end with one of the ten most frequent words, to the last ranks. For the smaller dataset the best results are achieved with the sole DRUID (see Table 8) and the frequency weighting does not seem to be beneficial, as highly frequent n-grams ending with stopwords are ranked higher in absence of POS filtering. This, however, is not observed for larger corpora. Here the best results for Medline are achieved with the frequency weighted DRUID. Whereas for French, the sole DRUID method performs best, the difference between the DRUID and the log-frequency-weighted DRUID is rather small. The low APs throughout can be explained by the large number of considered candidates. The second best scores are achieved with stop word based t-test (t-test + sw). C-value performs en par with frequency.

6.4 Components of DRUID

Here, we show different parameters for DRUID, relying on the English GENIA dataset without POS filtering of MWE candidates and by considering only terms with a frequency of 10 or more. Inspecting the two different components of the DRUID measure (see upper graph in Figure 2), we observe that the uniqueness measure contributes most to the DRUID score. The main effect of the incompleteness component is the downranking of a rather small number of terms with high uniqueness scores, which improves the overall ranking. We can also see that for the top ranked terms the negative incompleteness score does not improve over the frequency baseline but outperforms the frequency in the middle ranked candidates. Used in DRUID we observe a slight improvement for the complete ranking. We achieve a P@500 of 0.474 for the uniqueness scoring and 0.498 for the DRUID score.

When filtering similar entries, used for the uq scoring, by their similarity score (see bottom graph in Figure 2), we observe that the amount of similar n-grams considered seems to be more important then the quality of the similar entries: With the increasing filtering also the quality of extracted candidate MWEs diminishes.

7 Discussion and Data Analysis

The experiments confirm that our DRUID measure, either weighted with the MF or alone, works best across two languages and across different corpus sizes. It also achieves the best results in absence of POS filtering for candidate term extraction. The optimal weighting of DRUID depends on the nestedness of the MWEs: Using DRUID with the MF should be used when there are more than 20% of nested candidates and using the log-frequency or no frequency weighting when there are almost no nested candidate terms.

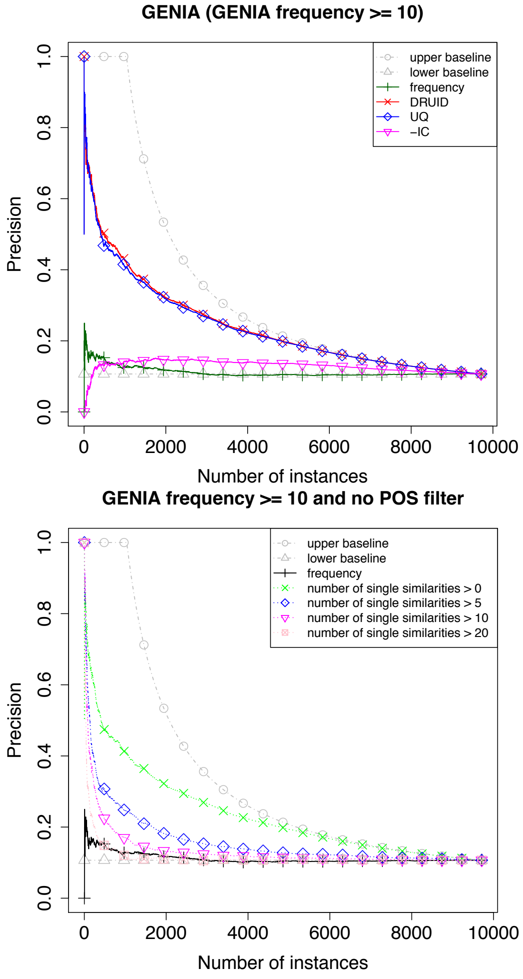

Figure 2: Results for the components of the DRUID measure (top) and for different filtering thresholds of the similar entries considered for the uniqueness scoring (bottom).

We show the best-ranked candidates obtained with our method and with the best competitive method in terms of P@100 for the two smaller corpora. Using the GENIA dataset, our log-frequency based DRUID (see left column in Table 9) ranks only true MWE within the 15 top-scored candidates.

The right-hand side shows results extracted with the pre-pruned MF method that yields three non-MWE terms. Whereas that could be a POS error,

log(freq)DRUID		**MF (pre-pruned)**	
NF-kappa B	1	T cells	1
transcription factors	1	NF-kappa B	1
transcription factor	1	transcription factors	1
I kappa B alpha	1	activated T cells	1
activated T cells	1	T lymphocytes	1
nuclear factor	1	human monocytes	1
human monocytes	1	I kappa B alpha	1
gene expression	1	nuclear factor	1
T lymphocytes	1	gene expression	1
NF-kappa B activation	1	NF-kappa B activation	1
binding sites	1	in patients	0
MHC class II	1	important role	0
tyrosine phosphorylation	1	binding sites	1
transcriptional activation	1	in B cells	0
nuclear extracts	1	transcriptional activation	1

Table 9: Top ranked candidates from the GENIA dataset using our method (left) and the competitive method (right). Each term is marked, whether the term is an MWE (1) or not (0).

the MF and also the C-value are not capable to remove terms starting with stop words. The DRUID score alleviates this problem with the uniqueness factor. For the French dataset our method ranks only one false candidate whereas the MF (post-pruned) ranks eight non-annotated candidates in the top 15 list (see Table 10).

DRUID		**MF (post-pruned)**	
hausse des prix	1	milliards de francs	0
mise en oeuvre	1	millions de francs	0
prise de participation	1	Etats - Unis	1
chiffre d' affaires	1	chiffre d' affaires	1
formation professionnelle	1	taux d' intérêt	1
population active	1	milliards de dollars	0
taux d' intérêt	1	millions de dollars	0
politique monétaire	1	Air France	1
Etats - Unis	1	% du capital	0
Réserve fédérale	1	milliard de francse	0
comit d' tablissement	1	directeur général	1
projet de loi	1	M. Jean	0
système européen	0	an dernier	1
conseil des ministres	1	années	1
Europe centrale	1	% par rapport	0

Table 10: Top ranked candidates from the SPMRL dataset for the best DRUID method (left) and the best competitive method (right). Each term is marked, if it is an MWE (1) or not (0).

Whereas the unweighted DRUID method scores better than its competitors on the large corpora, the best results are achieved when using DRUID with frequency-based weights on the smaller corpora. For a direct comparison we evaluated the small and large corpora using an equal candidate set. We observed that all methods computed on the large corpora achieve slightly inferior results than when computing them using the small corpora. Data analysis revealed that we would consider many of the high ranked ”false” candidates as MWE.

Therefore we extracted the top ten ranked terms, which are not annotated as MWE from the methods with the best P@100 performance, resulting to the log(freq) DRUID and the pre-pruned C-value methods.

First, we observed that the first ’false’ candidate for our method appears at rank 26 and at rank 1 for the C-value. Additionally, only ten out of the top 74 candidates are not annotated as MWEs for our method and 48 for the competitor. When searching the terms within the MeSH dictionary[10], we find seven terms ranked from our method and two for the competitive method.

8 Conclusion

Uniqueness is a new mechanism in MWE modeling. Whereas frequency and co-occurrence have been captured in many previous approaches (see Manning and Schütze (1999), Ramisch et al. (2012) and Korkontzelos (2010) for a survey), we boost multiword candidates t by their grade of distributional similarity with single word terms. We implement such contextual substitutability with a model where the term t can consist of multiword tokens and similarity is measured based on the right and neighboring word between all (single and multiword) terms. Since it is the default to express concepts with single words, a high uniqueness score is given to multiwords that belong to a category just as single words would. E.g. for an English open-domain corpus *hot dog* is most similar to the terms: *food*, *burger*, *hamburger*, *sausage* and *roadside*. Candidates with a low number of single word similarities also serve the same function, but more frequently we observe single n-grams with function words or modifying adjectives concatenated with content words, i.e. *small dog* is most similar to ”*various cat*”, ”*large amount of*”, ”*large dog*”, ”*certain dog*”, ”*dog*”. To be able to kick in, the measure requires a certain minimum frequency for candidates in order to find enough contextual overlap with other terms. Additionally, we also demonstrate effective performance on larger corpora and show its applicability when used in a complete unsupervised evaluation setting.

[10] http://www.nlm.nih.gov/mesh/

Acknowledgment

This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the context of the Software Campus project LiCoRes under grant No. 01IS12054 and the Deutsche Forschungsgemeinschaft (DFG) within the SeMSch project.

References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a French Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04)*, pages 2233–2236, Lisbon, Portugal.

Dimitra Anastasiou. 2010. *Idiom Treatment Experiments in Machine Translation*. Ph.D. thesis, Universität des Saarlandes, Saarbrücken, Germany.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling*, 1(1):55–95.

Stefan Bordag. 2008. A Comparison of Co-occurrence and Similarity Measures As Simulations of Context. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '14)*, pages 52–63, Haifa, Israel.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, pages 674–679, Istanbul, Turkey.

Guillaume Cabanac, Gilles Hubert, Mohand Boughanem, and Claude Chrisment. 2010. Tie-breaking Bias: Effect of an Uncontrolled Parameter on Information Retrieval Evaluation. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*, pages 112–123, Padua, Italy.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 515–521, Kyoto, Japan.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Stefan Evert. 2008. A lexicographic evaluation of German adjective-noun collocations. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE '08)*, pages 3–6, Marrakech, Morocco.

Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. 1998. The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98)*, pages 585–604, Heraklion, Greece.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14)*, pages 58–68, Baltimore, MA, USA.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 3.

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 (MWE '06)*, pages 12–19, Sydney, Australia.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–182.

Ioannis Korkontzelos. 2010. *Unsupervised Learning of Multiword Expressions*. Ph.D. thesis, University of York, UK.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. 2014. Yet Another Ranking Function for Automatic Multiword Term Extraction. In *Proceedings of the 9th International Conference on Natural Language Processing (PolTAL '14)*, pages 52–64, Warsaw, Poland.

Chris Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.

Hiroshi Nakagawa and Tatsunori Mori. 2002. A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*, COMPUTERM '02, pages 1–7, Taipei, Taiwan.

Hiroshi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proceedings of the Student Research Workshop of the 50th Meeting of the Association for Computational Linguistics (ACL Student Workshop '12)*, pages 1–6, Jeju Island, Korea.

Carlos Ramisch. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Ph.D. thesis, Universidade Federal Do Rio Grande do Sul.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL '14)*, pages 472–481, Gothenburg, Sweden.

Patrick Schone and Daniel Jurafsky. 2001. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP '01)*, pages 100–108, Pittsburgh, PA, USA.

Djamé Seddah, Marie Candito, Benoit Crabbé, and Enrique Henestroza Anguiano. 2012. Ubiquitous Usage of a Broad Coverage French Corpus: Processing the Est Republicain corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12)*, pages 3249–3254, Istanbul, Turkey.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, WA, USA.

Stefan Thater, Georgiana Dinu, and Manfred Pinkal. 2009. Ranking Paraphrases in Context. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer '09) in conjunction with the ACL '09*, pages 44–47, Suntec, Singapore.

Joachim Wermter and Udo Hahn. 2005. Effective grading of termhood in biomedical literature. In *Annual AMIA Symposium Proceedings*, pages 809–813, Washington D.C., USA.