Multiword Expression Identification
with Recurring Tree Fragments and Association Measures

Federico Sangati
Fondazione Bruno Kessler (FBK)
Trento, Italy
sangati@fbk.eu

Andreas van Cranenburgh
Huygens ING, Royal Netherlands Academy of Arts & Sciences; ILLC, Univ. of Amsterdam.
andreas.van.cranenburgh@huygens.knaw.nl

Abstract

We present a novel approach for the identification of multiword expressions (MWEs). The methodology extracts a large set of recurring syntactic fragments from a given treebank using a Tree-Kernel method. Differently from previous studies, the expressions underlying these fragments are arbitrarily long and can include intervening gaps. In the initial study we use these fragments to identify MWEs as a parsing task (in a supervised manner) as proposed by Green et al. (2011). Here we obtain a small improvement over previous results. In the second part, we compare various association measures in reranking the expressions underlying these fragments in an unsupervised fashion. We show how a newly defined measure (Log Inside Ratio) based on statistical parsing techniques is able to outperform classical association measures in the French data.

1 Introduction

According to many current linguistic theories, language users produce and understand sentences without necessarily fully decomposing them into ‘words’ and ‘rules’; rather, multiword units may function as the elementary building blocks (Goldberg, 1995; Kay and Fillmore, 1997; Stefanowitsch and Gries, 2003). A growing literature is emerging which focuses on “idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag et al., 2002) also referred to as multiword expressions (MWEs). These expressions, such as “to beat around the bush”, can be arbitrarily long. An important question for computational linguistics is how to identify such building blocks using statistical regularities in large corpora (Zuidema, 2006; Ramisch et al., 2012).

Most of the work on the identification of MWEs has focused on very short expressions, typically bigrams (Evert, 2005) or trigrams (Lyse and Andersen, 2012) using unsupervised techniques based on word association measures. Recent work (Green et al., 2011, 2013) has incorporated full phrase-structure trees in the process of multiword expression identification, obtaining a 36.4% F1 absolute improvement in MWE identification using a Tree-Substitution Grammar over an n-gram surface statistics baseline (Ramisch et al., 2010). However, one needs to note that the French Treebank (Abeillé et al., 2003) used in this study, contains explicitly tagged MWEs (as a special phrasal category), and therefore the comparison between supervised and unsupervised identification is not entirely fair.

In the current work, we present a hybrid method using both phrase-structure representation of MWEs, and association measures for ranking them in an unsupervised fashion (see table 1 for a quick comparison between the current work and previous approaches). We make use of a Tree-Kernel method (Collins and Duffy, 2002) for extracting a large set of recurring syntactic fragments from a given treebank.

The rest of the paper is organized as follows: in section 2 we present the idea of adopting recurring tree fragments extracted from a treebank using a Tree Kernel. In section 3 we introduce the treebanks from which tree fragments are extracted. Next we perform two types of experiments: in section 4 we employ the extracted fragments for supervised identification of multiword expressions as a supervised parsing task; in section 5, we compare how well different association measures rerank the expressions underlying the extracted fragments in an unsupervised fashion.
2 Recurring Fragments

In our work, we investigate ways of automatically detecting MWEs in large treebanks by searching for recurring patterns. The patterns consist of tree fragments that occur two or more times in the treebank. This is an ideal constraint if we want to assume that a necessary condition for a fragment to yield a MWE is to *recur multiple times in a representative corpus*.

This is also one of the original motivations behind the Data-Oriented Parsing (DOP) framework in which “idiomaticity is the rule rather than the exception” (Scha, 1990). For instance, if we have seen the MWE “pain in the neck” several times before, we should store the whole fragment for later use.

Data-Oriented Parsing has been most successfully implemented (Bod, 1992; Bod et al., 2003) with Tree Substitution Grammars (TSGs). A Tree-Substitution Grammar consists of a bag of elementary trees. In DOP, these are arbitrarily large fragments extracted from a treebank corresponding to syntactic constructions. They can include any number of lexical units, with possible intervening gaps, and are therefore very suited to represent MWEs ranging from fixed idiomatic cases such as “kick the bucket” to more flexible expressions such as “break X up” and “as far as X is concerned” to even longer constructions such as “everything you always wanted to know about X but were afraid to ask.”

Since extracting all possible fragments from a large treebank is impossible (the number of possible fragments grows exponentially with the size of a tree) it is necessary to work with a restricted set of fragments. Several sampling methods have been proposed (Bod, 2001; Zuidema, 2007; Cohn et al., 2010), but all include some limitations (e.g., use of random sampling methods, restriction in the size of the fragments, number of lexical items).

An alternative is to use a Tree Kernel which quantifies the similarity of trees (Collins and Duffy, 2002). Sangati et al. (2010) introduces FragmentSeeker, an algorithm based on a Tree Kernel that makes the similarities between trees explicit by extracting recurring tree fragments. FragmentSeeker is based on a dynamic programming algorithm which compares every pair of trees of a given treebank and extracts a list of maximal overlapping fragments in all the pairs.

In a recent effort, van Cranenburgh (2014) developed an improved algorithm\(^1\) for fragment extraction which runs in linear average time in the size of the treebank (it is 30 times faster than the original implementation on the Penn treebank). This substantial speedup is due to the incorporation of the Fast Tree Kernel (Moschitti, 2006), and opens up the possibility of handling much larger treebanks.

Figure 1 shows an example of a pair of trees sharing a common fragment (with lexical items depicted in blue and non-lexical terminals in green).

The fragments extracted with these tools have proven to be successful for several NLP tasks such as statistical parsing, as in DOP (Sangati and Zuidema, 2011; van Cranenburgh and Bod, 2013), authorship attribution (van Cranenburgh, 2012), and native language detection (Swanson and Charniak, 2012, 2013).

\(^1\)The tool is publicly available at https://github.com/andreasvc/disco-dop

<table>
<thead>
<tr>
<th></th>
<th>Ramisch et al. (2010)</th>
<th>Green et al. (2013)</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Association</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>measures</td>
<td>POS tags</td>
<td>flat rules</td>
<td>hierarchical</td>
</tr>
<tr>
<td>Syntax</td>
<td>flat rules</td>
<td>hierarchical</td>
<td></td>
</tr>
<tr>
<td>Gaps</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Representation</td>
<td>⟨JJ, mountain, NN_bike⟩</td>
<td>VP</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison of the current work with previous approaches.
3 Treebanks

We are using three different treebanks for extracting MWEs across three languages: French, Dutch and English. See table 2 for statistics on treebank sizes and number of fragments, and figure 2 for a comparison of the MWE annotations in the treebanks.

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Trees</th>
<th>Total Frags</th>
<th>Selected Frags</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>13K</td>
<td>274K</td>
<td>86K</td>
</tr>
<tr>
<td>Dutch</td>
<td>52K</td>
<td>536K</td>
<td>193K</td>
</tr>
<tr>
<td>English</td>
<td>500K</td>
<td>4.3M</td>
<td>2.8M</td>
</tr>
</tbody>
</table>

Table 2: Treebank size and number of fragments extracted and employed in the experiments. The last column reports the number of fragments after filtering out all those which do not contain at least a content word and a non-punctuation word.

3.1 French Treebank

We adopt the version of the French Treebank (Abeillé et al., 2003) from June 2010 used in Green et al. (2011). In this treebank MWEs are annotated with a flat bracketing (see figure 2a), that is, all words are grouped non-hierarchically, immediately under a single phrase which has a specific label per each phrasal category (e.g., MWV for verbal expression, MWN for nominal expressions, etc). We use this corpus for both supervised (parsing) and unsupervised (association measures) identification of MWEs.

3.2 Dutch Treebank

For Dutch, we employ the LASSY Small treebank (Noord, 2009) which is a syntactically annotated and manually verified corpus of 1 million words. As shown in figure 2b, the MWE annotation is flat as in the French Treebank, but a single category (MWU) is used to label them. We use this treebank for both supervised and unsupervised identification of MWEs.

3.3 Annotated English Gigaword

For English, we use the Annotated English Gigaword treebank\(^2\) which contains more than 180 million automatically parsed sentences. The size of this treebank is still prohibitively large even for the fast version of FragmentSeeker. We therefore decided to use only a sample of the treebank by selecting one out of every 150 sentences. This leaves us with a treebank of 500K structures, still 10 times larger than the Dutch treebank. However, since we want to extract MWEs, we are only interested in fragments with at least two lexical items. This restriction enables us to apply a further optimization to the algorithm which substantially boosts the extraction speed: after indexing sentences by the words they contain, we compare every tree structure only to other structures sharing at least two words.

\(^2\)http://catalog.ldc.upenn.edu/LDC2012T21
The annotation of the English Gigaword treebank follows the Penn Treebank scheme (Marcus et al., 1994) which does not include any special category for MWEs. As we have no gold standard for MWE annotation, we can only employ this treebank for unsupervised experiments and qualitative analysis. However, as shown in figure 2c, this annotation preserves the full hierarchical structure of MWEs and allows us to employ the full potential of Tree Kernels for extracting arbitrarily large MWEs with possible intervening gaps.

4 Finding MWEs by parsing

Green et al. (2011) introduce the idea of using a parsing model to identify MWEs. This is a supervised methodology as it requires a training treebank with gold MWE labels. The experiments of this section will therefore be performed on the French and Dutch treebanks.

4.1 Parsing Methodology

As parsing model we use the Double-DOP (2DOP) model (Sangati and Zuidema, 2011), as implemented in the disco-dop parser (van Cranenburgh and Bod, 2013). The resulting TSG grammar is constituted by the recurring fragments extracted from the training portion of the treebank (as explained in section 2) and additionally the Context-Free Grammar (CFG) rules occurring once (in order to ensure better coverage over the test sentences). In a TSG, fragments are combined by means of the substitution operation to derive the tree structures of novel sentences (see figure 4 for an example of fragments combination). We redirect the reader to Bod et al. (2003) for more details about TSG parsing.

In our models we use simple relative frequencies as fragment probabilities. As preprocessing we apply a set of manual state splits, heuristics for head-outward binarization, and an unknown word model for assigning POS tags to out-of-vocabulary words. For Dutch, we use the same preprocessing as described in van Cranenburgh and Bod (2013). For French, we apply similar preprocessing as Green et al. (2013).3

4.2 Results

In table 3 we present the comparison of the overall parsing results on the French and Dutch treebanks together with the MWE detection score. The overall parsing results (F1 score, exact match) are not specific to MWEs, but describe the general quality of the parsing model. The MWE-F1 score is an F1 score of correctly parsed MWE constituents.

For French we compare our model (2DOP) against two systems reported in Green et al. (2013), i.e., the factored Stanford parser and a TSG-DP parser in which tree fragments are drawn from a Dirichlet process (DP) prior (Cohn et al., 2010). Our system performs better than the other systems, both in terms of overall parsing results and MWE identification specifically.

For Dutch, since this is the first attempt to extract MWEs via parsing, we compare our result with a simple PCFG baseline. Our 2DOP model performs well above the baseline both in terms of parsing and MWE identification.

Finally, table 4 presents the detailed results for the identification of the MWEs for each category in the French treebank. Our system performs better in 4 out of 8 categories compared with the Stanford parser and the DP-TSG model. The Dutch results consist of a single category, so we do not report a further breakdown.

Table 3: Performance of the parsing models on the French and Dutch treebanks, with respect to parsing results (F1 score and exact match) and the MWE-F1 score, for sentences ≤40 words.

<table>
<thead>
<tr>
<th>Parser</th>
<th>F1</th>
<th>EX</th>
<th>MWE-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FRENCH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green et al. (2013): DP-TSG</td>
<td>76.9</td>
<td>16.0</td>
<td>71.3</td>
</tr>
<tr>
<td>Green et al. (2013): Stanford</td>
<td>79.0</td>
<td>17.6</td>
<td>70.5</td>
</tr>
<tr>
<td>disco-dop, 2DOP</td>
<td><strong>79.3</strong></td>
<td><strong>19.9</strong></td>
<td><strong>71.9</strong></td>
</tr>
<tr>
<td><strong>DUTCH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>disco-dop, PCFG baseline</td>
<td>63.9</td>
<td>21.8</td>
<td>50.4</td>
</tr>
<tr>
<td>disco-dop, 2DOP</td>
<td><strong>77.0</strong></td>
<td><strong>35.2</strong></td>
<td><strong>75.3</strong></td>
</tr>
</tbody>
</table>

Table 3: Performance of the parsing models on the French and Dutch treebanks, with respect to parsing results (F1 score and exact match) and the MWE-F1 score, for sentences ≤40 words.

5 Identifying MWEs with Tree Fragments and Association Measures

In this section we focus on the unsupervised detection of MWEs. We start with the same Tree Kernel
methodology illustrated in section 2 for extracting the set of recurring fragments from the various treebanks. Next, we apply various association measures (AMs) for ranking these fragments and compare how they perform in distinguishing those fragments underlying MWEs from the others. In section 5.3 we conduct a case study on the English treebank for which we have no MWE annotations, whereas in section 5.4 we apply a quantitative analysis to assess how the AMs perform in the French and the Dutch treebank (for which we have gold MWE annotations).

5.1 Signatures

Differently from most existing works on MWEs discovery, our methodology does not focus on MWEs of a specific type or size. However, the association measures that are commonly employed are strongly influenced by the length of the expressions, i.e., shorter expressions tend to have higher association scores. Moreover, since we also take into account fragments with possible gaps, we need to be careful in distinguishing fully lexicalized expressions from those containing intervening phrasal categories.

We therefore devise a way to partition the set of extracted fragments into a number of bins. All fragments belonging to the same bin share the same signature and are therefore mutually comparable (in terms of their association scores). The signature of a fragment is a sequence \( L, X \) of symbols obtained by mapping each frontier node of the fragment to \( L \) if it is a lexical node, or \( X \) if it is a non-lexical node. Figure 3 shows an example of a fragment and its corresponding signature.

5.2 Association Measures

A number of Association Measures (AM) have been defined in the literature to assess the cohesiveness of a potential MWE. In this work we take into consideration two standard association measures, the Pointwise Mutual Information (PMI) and the Log-Likelihood Ratio (LLR). Both AMs are generalized to arbitrarily long expressions, and are defined over the sequence of symbols \( S_1, S_2, \ldots, S_n \), where \( S_i \) is the pair \((\text{pos}_i, \text{word}_i)\), with \( \text{pos}_i \) and \( \text{word}_i \) being the pre-terminal label and lexical item of the \( i \)-th frontier node, respectively; \( \text{word}_i = \emptyset \) if the \( i \)-th frontier node is a non-lexical item. In addition, we define a novel association measure, namely the Log Inside Ratio (LIR), based on probabilities of a probabilistic TSG underlying the extracted fragments.

PMI The Multivariate Generalization of Pointwise Mutual Information, also referred to as Total Correlation (Watanabe, 1960) and Multi-Information (Studený and Vejnarová, 1998; Van de Cruys, 2011), is defined as follows:

\[
\text{PMI}(S_1, S_2, \ldots, S_n) = \log \frac{p(S_1, S_2, \ldots, S_n)}{\prod_{i=1}^{n} p(S_i)}
\]

where \( p(S_1, S_2, \ldots, S_n) \) is the relative frequency with which the signature \( S_1, S_2, \ldots, S_n \) has been seen within the set of fragments sharing the same signature, and \( p(S_i) \) is the relative frequency of seeing the symbol \( S_i \) in the \( i \)-th position of the signature within the same set of fragments.

LLR The Log-Likelihood Ratio generalized for a sequence with an arbitrary number of symbols (Su, 1991) is defined as follows:

\[
\text{LLR}(S_1, \ldots, S_n) = \log \frac{p(S_1, \ldots, S_n)}{\sum_{\sigma \in \text{CSP}(S_1, \ldots, S_n)} \prod_{s \in \sigma} p(s)}
\]

where the numerator is as in PMI, while the denominator represents the probability of the sequence to...
be derived from contiguous spans. More precisely, \( \text{CSP}(S_1, \ldots, S_n) \) returns the ways (\( \sigma \)) of partitioning the sequence \( S_1, \ldots, S_n \) in contiguous spans (\( s \)).\(^4\)

**LIR** The Log Inside Ratio is a newly derived association measure which specifies the probability that a Probabilistic TSG (PTSG) grammar generates a given fragment in a single step with respect to the total probability of generating it in any possible way, i.e., by combining smaller fragments together. Figure 4 shows an example of how a TSG can generate the same fragment in multiple ways. The LIR is computed as follows:

\[
\text{LIR}(\text{frag}) = \log \frac{p(\text{frag})}{\text{inside}(\text{frag})}
\]

where the numerator is the probability of the fragment according to the PTSG extracted from the treebank,\(^5\) while the denominator is the total probability with which the grammar generates the given fragment starting from its root category (in any possible way).

![Figure 4: Example of how a TSG can generate the same fragment in two different ways, i.e., in a single step (above), and in 3 subsequent steps (below).](image)

In this initial study we limited the qualitative analysis to the PMI association measure.

The histogram in figure 5 reports the distribution of the extracted fragments in the most common signature bins. This includes fragments with up to 7 terminals at the frontier nodes, with at most 3 non-lexical nodes (\( X \) in the signatures). Tables 5 and 6 present a list of fragments starting with the verb *take* with and without a gap in the second position, sorted by the PMI measure. In both cases there is a contrast between MWEs at the top of the list (e.g., *take into account*) and more compositional expressions at the bottom (e.g., *take QP years to, take the money*).

![Figure 5: Distribution of the 2.8M recurring fragments extracted from the English treebank into the various signature bins. Only bins with at least 100 fragment types are reported.](image)

**5.4 Quantitative Results on French and Dutch**

Quantitative evaluation of MWE identification is a non-trivial task. Typically, association measures are tuned so that only expressions above a specific threshold are considered MWEs. Alternatively, precision and recall measures on a full reference data or on n-best lists are used (Evert and Krenn, 2001). In our case the task is more challenging as we would need to fix a different threshold value for each set of fragments sharing the same signature. We therefore decided to resort to a novel evaluation metric which would enable us to compare how the various AMs rerank the full list of expressions sharing the same signature in a more neutral and informative way.

We do so by calculating, for each signature bin, the percentage of MWEs present in subsequently smaller portions of the reranked list, limiting the evaluation...
to fewer and fewer candidates at the beginning of the list (as association measures tend to place MWEs on top). This metric is similar to the “precision at \( k \)" used in Information Retrieval, except that instead of using a fixed integer \( k \), we use varying portions of the list (i.e., \( 1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{10} \)).

Figure 6 shows the resulting graphs for the three AMs and the most common signatures in the French and Dutch treebanks. All curves are usually monotonically increasing, indicating that for all measures the concentration of MWEs increases at the top of the reranked list. PMI and LLR often overlap (they are mathematically identical for expressions of length 2), with LLR being slightly better for French and PMI for Dutch. Finally LIR is consistently better than the other 2 AMs for French while being worse or on a par with the others for Dutch. We are currently investigating the reason for this discrepancy. Our current hypotheses are: (i) the French treebank makes use of several MWE categories while the Dutch treebank has a single MWE category, and (ii) Dutch MWEs tend to be less rigid than the French ones.

Table 7 shows a single-figure F1 evaluation of the three AMs, obtained by aggregating the top 1/5 candidates of each bin. For this evaluation, recall and precision are computed, with the gold set consisting of all the extracted lexicalized fragments with MWE gold tags. According to these results the Log Inside Ratio (LIR) performs best for both French and Dutch. This evaluation is not ideal, as our method aims to go beyond the small, contiguous MWE strings annotated in the treebanks. In addition, manual inspection of the selected candidates reveals that many of them are MWEs, while not part of the gold standard. This should be addressed in future work with a manual evaluation.

<table>
<thead>
<tr>
<th>Treebank</th>
<th>PMI</th>
<th>LLR</th>
<th>LIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>33.0</td>
<td>32.3</td>
<td>45.8</td>
</tr>
<tr>
<td>Dutch</td>
<td>49.4</td>
<td>46.6</td>
<td>50.5</td>
</tr>
</tbody>
</table>

Table 7: F1 scores for the top 1/5 candidates of each bin as ranked by the three AMs evaluated against MWEs in extracted recurring fragments.

6 Conclusion

We have presented a novel approach for the identification of MWEs based on recurring fragments automatically extracted from a treebank. We have shown that a probabilistic tree-substitution grammar (PTSG) constructed with these fragments outperforms previous results for the supervised identification of MWEs. Finally we have conducted a study to assess how various association measures (AMs) can rerank the extracted fragments for the unsupervised identification of MWE. Here we proposed a new measure based on PTSG, the Log Inside Ratio, which shows competitive results when compared against other classical association measures.

Acknowledgments

We kindly acknowledge the three anonymous reviewers and Katja Abramova for very useful feedback as well as the PARSEME European Cost Action (IC1207) for promoting this collaboration.
Figure 6: Results for the French and Dutch treebanks when ranking of the MWEs for various signatures according to several association measures. Each line reports how the percentage of MWEs (y-axis) changes when restricting the list to fewer and fewer top candidates. More specifically, we compute the percentage of MWE in the full list of fragments (1), in the first half (1/2), the first third (1/3), and so on until the first tenth (1/10).
References


