Identification of Idiomatic Expressions Using Parallel Subtitle Corpora

Nina Viereckel, Jörg Tiedemann

Uppsala University
Department of Linguistics and Philology
ninaviereckel@gmail.com, jorg.tiedemann@lingfil.uu.se

1. Introduction

This paper deals with the automatic identification and extraction of idiomatic expressions from parallel corpora. Idiomatic expressions, a subset of multiword expressions (Sag et al., 2001), henceforth called MWEs, are lexical items consisting of multiple simplex words that are generally not fully compositional and therefore problematic to analyse and process in applications related to natural language processing. In the past two decades, there has been a growing interest in the automatic identification and extraction of idiomatic expressions and other kinds of MWEs. Among the numerous approaches to automatically extract these expressions from text, it has been shown that the use of parallel corpora delivers satisfying results.

In this work, we use statistical association measures to extract idiomatic expressions and improve the resulting ranking by using the alignment information provided by parallel corpora. This approach is based on the work of Villada Moirón and Tiedemann (2006). In contrast to their approach, which was done on Dutch, we will perform our experiments on English. In addition, we expand the set of MWE candidates by adding other structures than VERB PP to the set of extracted idioms, and we will test the method on a different corpus, the OpenSubtitles2012 dataset, a collection of TV series and movie subtitles, compiled from the website OpenSubtitles (http://www.opensubtitles.org/).

2. Related Work

This work draws heavily on the work of Villada Moirón and Tiedemann (2006). They propose an approach to extract Dutch idiomatic expressions from Europarl, using association measures such as log-likelihood and salience as well as a head dependence heuristic. They assume that an expression has a non-compositional meaning if its translation is not a combination of its components' translations, and that an automatic word aligner would produce a larger variety of links when encountering non-compositional expressions. To validate this hypothesis, they take the top 200 candidates and rerank them according to two other measures, the proportion of default alignments among the links found for MWE components, and translational entropy, which are both computed using the word alignment between source and target language in the parallel corpus. With these methods, they are able to achieve up to 93.2% and 91.7% uninterpolated average precision, respectively, on Dutch to German word alignments, compared to the 75.5% baseline using only association measures.

3. Methodology

Preprocessing: We are working with the German-English portion of the OpenSubtitles2012 corpus (Tiedemann, 2012), and Europarl (Koehn, 2005) for comparison. Automatically aligned and truecased versions of the corpora are available from the OPUS website (http://opus.lingfil.uu.se/). We use TreeTagger (Schmid, 1994) to do part-of-speech tagging, as it provides both tag and lemma information for a given token, and MaltParser (Nivre et al., 2006) to parse the corpora. For parsing, a pre-trained model for English, trained on the Penn Treebank, is available for use from the MaltParser homepage (http://www.maltparser.org).

MaltParser requires the data to be converted into the CoNLL format, so we convert our previously tagged and lemmatised datasets by utilising the POS tags for both required POS tag fields and the lemmatisation information for the LEMMA field. After parsing, we get a new CoNLL file that also contains the head of the token and the dependency relation to it. With this information, we can continue with the extraction of idiomatic expressions.

Extraction: The following support verbs are selected to extract idiomatic expressions: *come*, *go*, *take*, *give*, *hit*, *throw*, *rise*, *fall*, *do*, *make*, *stand*, *put*, *bring*, *stay*, and *hold*. According to Butt (2010), the first 10 verbs are common examples for support verbs that are used crosslinguistically. Additionally, we choose five other common verbs that can function as support verbs.

We extract n-grams from the corpus as VERB NP and VERB PP tuples. The extraction is done by collecting syntactic n-grams, fragments of dependency trees, extracted using the dependency relations generated by MaltParser. Since we are only interested in the statistics of VERB NP and VERB PP tuples, we extract all n-grams containing a verb and a nominal phrase, or a verb, a nominal phrase and prepositional phrase.

The subtitle data contains some noise, thus we only consider tuples that occur at least 5 times. For each tuple, we compute log-likelihood scores with UCS (http://www.collocations.de/software.html), a tool that calculates association measures (Evert, 2005). We also compute salience scores (Kilgarriff and Tugwell, 2002). Furthermore, we measure the head dependence of each tuple by computing the observed entropy between its PP or NP, respectively, and the different verbs it can occur with (Merlo and Leybold, 2001). After ranking the candidates by uniformly combining the ranks assigned to each tuple, we select the top 200 candidates for further processing.

Expression	Local links	Global links
go	go: 4	gehen: 24537
	ist: 3	los: 14876
	NO_LINK: 2	geh: 8340
to	to: 8	zu: 136614
	in: 4	,: 93506
	zum: 3	mit: 28887
rehab	Entzug: 4	Reha: 43
	rehab: 4	der: 32
	die: 3	Entzug: 29

Table 1: Excerpt of the link lexica for ”go to rehab”.

Alignment Collection and Reranking: We collect two kinds of links from the different word alignments for the German-English portion of each corpus. First, we create a global link lexicon which contains all the occurring translations that a word is linked to in the dataset. Then, for each candidate expression, we collect alignment links within the context of its tuple, meaning that we only collect the possible translations for each word in the tuple if they occur in the context of the expression. If a word is not linked to another word in the source language, we add NO_LINK to the local link lexicon. An excerpt of the link lexica can be seen in Table 1.

To rerank our candidates, we compute the translational entropy (Melamed, 1997) of each expression by utilising the previously compiled local link lexicon. For each component of an expression, the entropy of its aligned target words is calculated as follows.

$$H(T_s|s) = -\sum_{t\in T_s} P(t|s)\log_2 P(t|s) \qquad (1)$$

Finally, we take the average translational entropy of an expression's components to assign a score to our candidates.

We also measure the proportion of default alignments (Villada Moirón and Tiedemann, 2006). This measure takes into account that the default alignments (i.e., the four most commonly aligned translations in the global link lexicon) of the components of an idiomatic expression will differ from the links in its local link lexicon. It is calculated as:

$$pda(S) = \frac{\sum_{s\in S}\sum_{d\in D_S} align\_freq(s,d)}{\sum_{s\in S}\sum_{t\in T_S} align\_freq(s,t)} \qquad (2)$$

where $align\_freq(s,d)$ is the number of times a word s is linked to one of the default alignments D_s and $align\_freq(s,t)$ is the alignment frequency of word s to word t in the context of an expression S.

We experiment with the following symmetrised alignment types: *source-to-target* (srctotgt), *target-to-source* (tgttosrc), *intersect*, and *grow-diag-final-and* (g-d-f-a).

4. Evaluation and Results

We use uninterpolated average precision (Manning and Schütze, 1999), henceforth abbreviated as *uap*, to assign a score to each of our rankings.

The results can be seen in Tables 2 and 3. The method performs better for OpenSubtitles than for Europarl on VERB NP tuples, while it performs worse on VERB PP expressions. This is mostly due to the fact that the initially extracted candidates contain more than 40 instances of expressions starting with *come on* (*”come on, mate”*) which were incorrectly parsed as being VERB PP candidates, displaying the weaknesses of using a rather noisy dataset. In all cases we see a considerable improvement over the baseline which consists of using only association measures and the head dependence heuristic. The advantages of using a subtitles corpus become obvious when looking at the extracted expressions: We are able to identify rather colloquial examples, such as *go to hell*, *give a damn*, *take a leak*, or *hit the fan*.

Subtitles	**srctotgt**	**tgttosrc**	**intersect**	**g-d-f-a**
pda	0.626	0.633	0.616	0.612
entropy	0.622	0.689	0.479	0.694
baseline	0.420	0.420	0.420	0.420
Europarl	**srctotgt**	**tgttosrc**	**intersect**	**g-d-f-a**
pda	0.743	0.791	0.754	0.761
entropy	0.738	0.790	0.647	0.757
baseline	0.712	0.712	0.712	0.712

Table 2: Results for VERB PP extraction (*uap*).

Subtitles	**srctotgt**	**tgttosrc**	**intersect**	**g-d-f-a**
pda	0.860	0.867	0.895	0.863
entropy	0.814	0.892	0.775	0.886
baseline	0.826	0.826	0.826	0.826
Europarl	**srctotgt**	**tgttosrc**	**intersect**	**g-d-f-a**
pda	0.835	0.864	0.848	0.845
entropy	0.803	0.851	0.783	0.840
baseline	0.769	0.769	0.769	0.769

Table 3: Results for VERB NP extraction (*uap*).

5. Conclusions

We have presented the results of applying the approach of Villada Moirón and Tiedemann (2006) to a subtitles corpus and expanding the set of candidates by adding VERB NP tuples to the set of extracted idioms. For both types of constructions we obtain encouraging results, confirming the findings of Villada Moirón and Tiedemann (2006).

In future work, we will apply this approach to other languages, such as German or Swedish, and use word alignments to several languages, in particular non-germanic ones. Furthermore, we will investigate if grouping the subtitles by genre and date information will yield interesting results.

References

Miriam Butt. The light verb jungle: Still hacking away. In *Complex Predicates*. Cambridge University Press, 2010.

Stefan Evert. The statistics of word cooccurrences: Word pairs and collocations, Ph.D. Thesis, 2005.

Adam Kilgarriff and David Tugwell. Sketching words. In *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, page 125–137, 2002.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, page 79–86. AAMT, 2005.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

I. Dan Melamed. Measuring semantic entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, page 41–46, 1997.

Paola Merlo and Matthias Leybold. Automatic distinction of arguments and modifiers: The case of prepositional phrases. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL–2001)*, page 121–128, 2001.

Joakim Nivre, Johan Hall, and Jens Nilsson. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC 2006*, page 2216–2219, 2006.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing 2002*, page 1–15, 2002.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. 1994.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC 2012*. European Language Resources Association (ELRA), 2012.

Begoña Villada Moirón and Jörg Tiedemann. Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, 2006.