Determining the Multiword Expression Inventory of a Surprise Language

Bahar Salehi,[1] **Paul Cook**[2] and **Timothy Baldwin**[1]

[1] NICTA Victoria Research Laboratory
Department of Computing and Information Systems
The University of Melbourne, Victoria 3010, Australia

[2] Faculty of Computer Science
University of New Brunswick
Fredericton, NB E3B 5A3, Canada

bsalehi@student.unimelb.edu.au, paul.cook@unb.ca, tb@ldwin.net

Abstract

Much previous research on multiword expressions (MWEs) has focused on the token- and type-level tasks of MWE identification and extraction, respectively. Such studies typically target known prevalent MWE types in a given language. This paper describes the first attempt to learn the MWE inventory of a "surprise" language for which we have no explicit prior knowledge of MWE patterns, certainly no annotated MWE data, and not even a parallel corpus. Our proposed model is trained on a treebank with MWE relations of a source language, and can be applied to the monolingual corpus of the surprise language to identify its MWE construction types.

1 Introduction

Multiword expressions ("MWEs") are word combinations which have idiosyncratic properties relative to their component words (Sag et al., 2002; Baldwin and Kim, 2010), such as *taken aback* or *red tape*. The need for an explicit model of MWEs has been shown to be important in NLP tasks including machine translation (Venkatapathy and Joshi, 2006), parsing (Constant et al., 2012), and keyphrase/index term extraction (Newman et al., 2012). However, existing approaches to MWE identification/extraction typically target specific MWE types that are known to be prevalent in a given language, such as: (a) compound nouns in languages such as English (Copestake, 2003; Ó Séaghdha, 2008), German (Schulte im Walde et al., 2013) and Japanese (Tanaka and Baldwin, 2003); (b) light verb constructions (LVCs) in languages such as English (Butt, 2003), Persian (Karimi-Doostan, 1997) and Italian (Alba-Salas, 2002); and (c) compound verbs in languages such as Japanese (Uchiyama et al., 2005). Note here that the combination of highly-productive MWE types can vary greatly across languages: English is rich with compound nouns and LVCs are also common, but lacks compound verbs; Persian is rich with LVCs and adjective–noun compounds, but has very few compound nouns and compound verbs; and Japanese is rich with LVCs and compound nouns and verbs, but adjective–noun MWEs are rarer. Even for collocation extraction, this knowledge is generally assumed for a given language, in targeting only highly productive constructions such as adjective–noun or verb–noun collocations (Krenn and Evert, 2001; Pecina, 2008).

But what if the language of interest is one where no such prior knowledge exists, e.g. because it is a "surprise" language where rapid deployment of language technologies is required and there is no access to an informant with sufficient linguistic training to be able to inventorise the MWE types in the language (Oard, 2003; Maynard et al., 2003)? Here, there is little expectation of success without an automatic method for determining the inventory and relative frequency of MWEs in a given language. This provides the motivation for this paper: can we develop a method for automatically profiling the MWE inventory of a novel language based simply on a monolingual corpus of that language, and a treebank in a second language such as English?

We carry out this research in the Universal Dependency ("UD") framework (Nivre et al., 2016), using the method of Duong et al. (2015) to induce a delexicalised dependency parser for the surprise language, based on a supervised parsing model for a language such as English where we have a well-developed treebank in the UD. Given the parser output over a monolingual corpus in the surprise language, we then apply one of two methods to extract our MWE profile: (1) a baseline method, where we simply

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 471–481, Osaka, Japan, December 11-17 2016.

extract out delexicalised dependency tuples of relation type `mwe` or `compound` (including the POS tags), aggregate the counts of the pos–relation–pos triples, and extract the most frequent triples; and (2) a supervised reranker over the delexicalised dependency tuples, to better deal with noise in the output of the delexicalised dependency parser.

One additional contribution of this paper is analysis of MWE annotation across different languages in the UD. We find that there are a number of competing styles of annotation, and very different levels of thoroughness in the annotation of MWEs. As part of this, we perform an "oracle" analysis of MWE extraction based on the gold-standard treebank annotations for a given language, and find that the results vary greatly between languages, due to annotation divergences. Using the supervised reranking method, however, and incorporating more and more languages for training (but holding out the surprise language), we find that we are able to "smooth" annotation differences between languages.

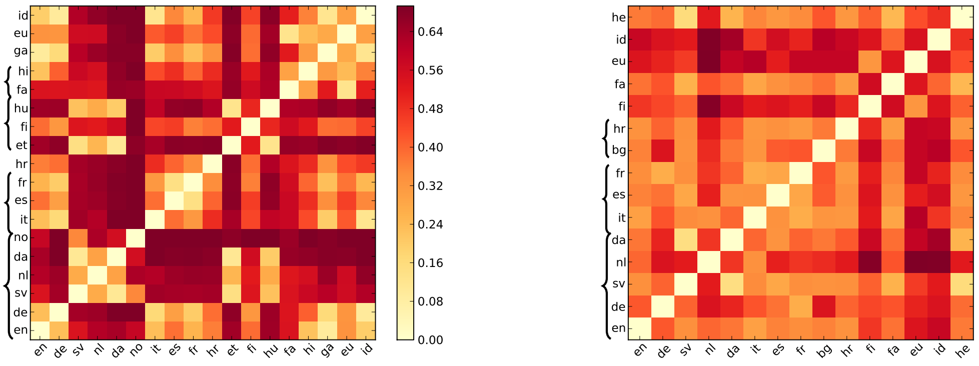
2 Related Work

There is a wealth of research on MWE identification (i.e. distinguishing MWEs from non-idiosyncratic combinations at the token level) and extraction (i.e. determining at the type level which word combinations in a corpus are MWEs). Many of these methods are customised to particular MWE constructions which are known to exist in a given corpus, e.g. noun compounds (Lapata and Lascarides, 2003; Tanaka and Baldwin, 2003), verb particle constructions ("VPCs": Baldwin and Villavicencio (2002; Baldwin (2005)), determinerless prepositional phrases (Baldwin et al., 2004; van der Beek, 2005), or compound verbs (Breen and Baldwin, 2009). There is also a significant body of work on general-purpose MWE extraction, often based on statistical association measures applied to either a monolingual corpus (Evert and Krenn, 2005; Pecina, 2008; Ramisch, 2012) or a parallel corpus (Melamed, 1997; Moirón and Tiedemann, 2006). Even here, however, POS-based constraints are generally applied on the types of MWE that are extracted (e.g. noun–noun or verb–noun bigrams). There are also methods for identifying MWEs in context using supervised models (Diab and Bhutada, 2009; Li and Sporleder, 2010; Schneider et al., 2014), which require exhaustive annotation of MWE token occurrences in a corpus. All of this research differs from our work in that it either assumes knowledge of the type(s) of MWE to extract for a given language, or requires explicitly annotated MWE data in that language.

Closer to home, there has recently been work on general-purpose, unsupervised approaches to MWE extraction, making no assumptions about the types of MWE that exist in a given language (Newman et al., 2012; Brooke et al., 2014). Here, however, the definition of MWE tends to be blurred somewhat to focus on index terms or "formulaic language", i.e. idiomatic expressions with statistically-marked properties in a given corpus — blurred in the sense that many MWEs are not statistically marked, and also that they include formulaic expressions such as *in this paper* that are not formally MWEs.

Also related is recent work on resource development for low-resource languages, such as dependency parsing based on transfer learning from a higher-density language (Naseem et al., 2012; Täckström et al., 2013; Duong et al., 2015). For example, Duong et al. (2015) proposed a neural network-based parser that transfers dependency relations across languages without requiring a parallel corpus. They learn syntactic cross-lingual word embeddings by training the skip-gram model (Mikolov et al., 2013) on a representation of the original text in which the context of each token is represented by its universal POS tags (Petrov et al., 2012). They then incorporate these word embeddings in a transition-based neural network dependency parser (Chen and Manning, 2014).

Our proposed method is the first attempt to learn the MWE profile of a language with no knowledge of the target language except for POS tags (which themselves can be induced automatically, with little or no annotated data: Garrette and Baldridge (2013), Duong et al. (2014)), and no parallel corpus. We train a delexicalised dependency parser based on transfer learning (involving no syntactic annotations for the target language), and train a reranking model based on observed MWEs in only the source language(s).



(a) `compound`

(b) `mwe`

Figure 1: Cross-lingual similarity of MWE pattern distributions using JSD

3 Resources

The Universal Dependency Treebank[1] ("UD") is a universal annotation scheme for dependency parsing that is consistent among languages with the goal of cross-lingual learning (Nivre et al., 2016). It is made up of a universal part-of-speech (POS) tag set (Petrov et al., 2012) and universal dependency relation set, and a set of treebanks.

In this paper, we use v1.2 of UD. MWEs are labeled as either `name` (for named entities), `compound` (for binary compound expressions) or `mwe` (for fixed expressions). In this work, we focus specifically on `mwe` and `compound`, as named entity recognition is a specialised subtask of MWE identification with its own dedicated literature (Maynard et al., 2003; Huang et al., 2003; Steinberger and Pouliquen, 2007), and there is every expectation that all languages contain named entities. Although the documentation for UD provides definitions of how to distinguish `mwe` and `compound` in labelling MWEs, there seem to be major inconsistencies in how they have been interpreted for particular languages: some languages do not use these relations at all, while others only annotate a subset of MWE types with these relations.

The languages examined in this paper are as follows, in descending order of prevalence[2] of MWE annotations in UD (as indicated in parentheses):

> Hindi (14.0%), **Indonesian** (9.2%), **Persian** (7.5%), **Croatian** (6.7%), **English** (6.2%), **Swedish** (5.4%), Estonian (4.9%), Irish (4.7%), Finnish (4.4%), Basque (3.7%), Hungarian (2.7%), Dutch (2.2%), Norwegian (1.6%), Danish (1.5%), French (1.5%), Italian (0.8%), Spanish (0.8%), Hebrew (0.7%), Bulgarian (0.6%), German (0.4%)

These were selected based on the fact that they have at least 100 individual occurrences of the `mwe` or `compound` relation. The 5 languages in bold were selected as our test languages, based on the high prevalence of MWE annotations and diversity of MWE patterns.[3] Here and for the remainder of the paper, we define "MWE pattern" to be an ordered tuple of the form $\langle \text{pos}_h, \mathsf{rel}, \text{pos}_d \rangle$, where pos_h is the POS of the head, and pos_d is the POS of the dependent in the triple. Based on this definition, English has 56 distinct MWE patterns, Croatian 49, Persian 48, Swedish 45, and Indonesian 26.

4 MWE Patterns

This paper investigates the profile of MWE patterns in a given language, in the form of delexicalised dependency tuples. The most frequent patterns in our 5 target languages with the `compound` relation

[1] `https://universaldependencies.github.io/docs/`

[2] the proportion of MWE tokens

[3] We discarded Hindi despite the high proportion of MWEs because: (1) it only covered `compound` relations, and has no `mwe` relations, and (2) it has a low number of distinct MWE patterns (23), and as such appeared skewed in its annotation.

are: NOUN–NOUN; PROPN–PROPN (i.e. proper noun dependencies, which should be annotated with the name relation rather than compound, according to the annotation guidelines); and VERB–NOUN, which includes LVCs. There are also other noticeable patterns such as VERB–ADV(erb) and VERB–ADP(osition), corresponding to VPCs (Schulte im Walde, 2004; Baldwin, 2005).

mwe patterns are more diverse than compound patterns: compound patterns mostly involve nouns and verbs, while mwe patterns involve a diverse range of POS types, such as ADP–ADP or ADV–ADV, and pairings including CONJ(unctions) or SCONJ (subordinating conjunctions).

We additionally measured the similarity between the MWE pattern probability distribution of the different languages using Jensen–Shannon divergence, as shown in Figure 1 for all languages in UD. To make comparison between related languages easier, we clustered the languages by language family. According to Figure 1, there is no clear indication that languages of the same family have similar MWE patterns, which is something that we might have expected.

These results suggest that although the ultimate goal of the UD project is to have compatible annotations, the MWE annotations are not, at present, consistent. In fact, annotation divergences would appear to be more noticeable than linguistic differences. For example, the Norwegian treebank annotates only VPCs (and not multiword compound nouns, e.g.), and the mwe relation is not used at all. That is, the observed differences in MWE patterns certainly reflect differences between languages, but greater than this, they capture differences in the annotation process between different languages.

We also examined the annotation consistency of MWEs between the Train+Dev sets and Test set of each language (based on the provided splits), and observed high consistency (low JSD) between the existing patterns in these sets for the same language. The JSD on compound patterns are all below 0.10, except for Spanish (0.23) and French (0.18). Due to the diversity of mwe patterns, the JSD is less consistent within each language, with Croatian (0.64) and German (0.44) being notably high, and the rest of the languages below 0.25. This shows that annotation is quite consistent within each language.

Therefore, despite the cross-lingual annotation inconsistency, our corpora appear to be internally consistent enough to train a model over, based on the observed MWEs in a language.

5 Methodology

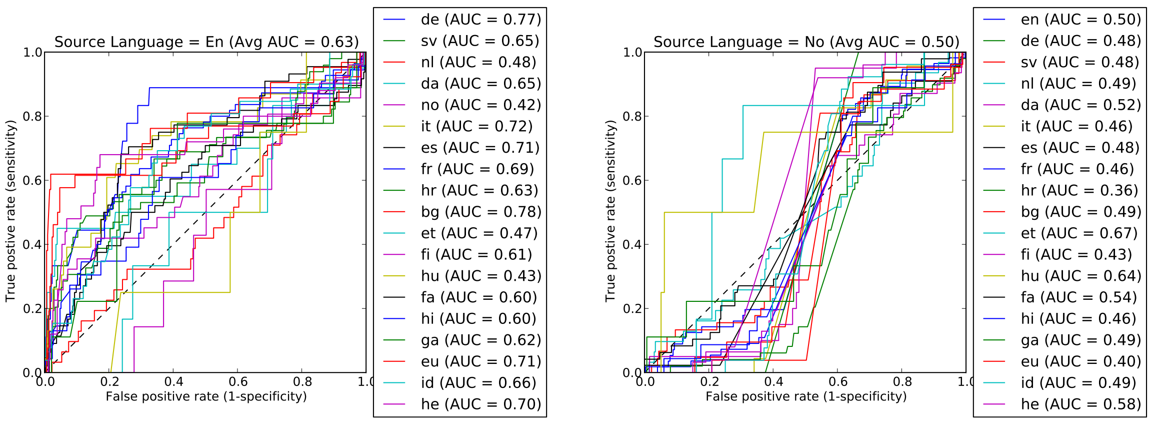
In this work, we measure the likelihood of the triple $\langle \text{pos}_h, \mathsf{rel}, \text{pos}_d \rangle$ being an MWE pattern in the target language. The scores are measured according to the respective lexical instances of each triple in the source language, aggregated to compute scores for each triple, and used to train a support vector regression (SVR: Joachims (2006)) model.

The gold-standard labels to train the model are based on the dependency relations: the value is set to 1 if the dependency is compound or mwe, and 0, otherwise.

We use 8 features in our proposed method: pointwise mutual information (PMI), ϕ-square, the Dice coefficient, student's t test, log-likelihood ratio, pattern fixedness, token/type ratio for a given triple, and token/token ratio across all relations.

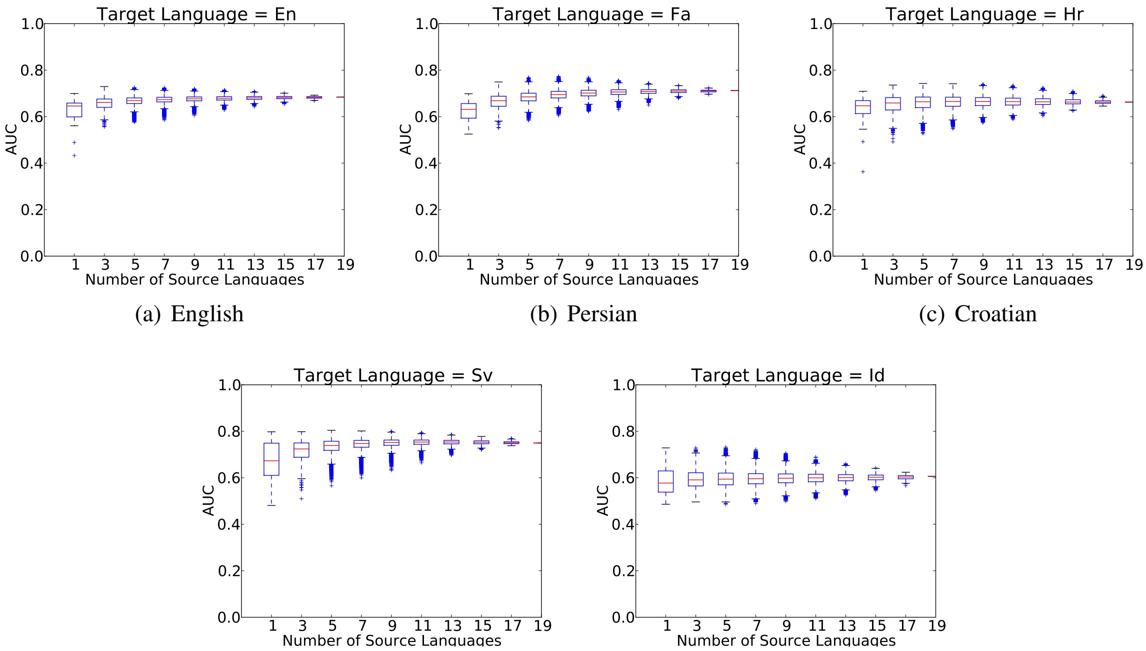
$$\begin{aligned}
\text{PMI}(x,y) &= \log \frac{p(x,y)}{p(x)p(y)} \\
\phi^2 &= \frac{(n(x,y)n(\bar{x},\bar{y}) - n(x,\bar{y})n(\bar{x},y))^2}{n(x)n(\bar{x})n(\bar{y})n(y)} \\
\text{Dice coefficient} &= \frac{2n(x,y)}{n(x)+n(y)} \\
\text{t}(x,y) &= \frac{\frac{n(x,y)-(n(x)*n(y))}{\text{total words}}}{\sqrt{n(x,y)}}
\end{aligned}$$

where $n(.)$ is the number of occurrences, and $\bar{x}$ is the number of all instances except x. Pattern fixedness is measured via entropy as $\text{H}(\Pr(D(x,y)))$, where $D(x,y)$ is the difference between the linear position of the head and dependent, binned as follows: $\textit{posdiff} \in \{(-\infty,-2), -2, -1, 1, 2, (2,\infty)\}$.



(a) English (b) Norwegian

Figure 2: Selecting a language as a source language



(a) English (b) Persian (c) Croatian

(d) Swedish (e) Indonesian

Figure 3: Distribution of ROC AUC scores when combining source languages

These features are first computed for each lexical instance of a given pattern, and then aggregated to calculate the overall feature values for each triple, using either: (a) the mean (in the case of pattern fixedness); or (b) the median (in the case of the other measures).[4]

After computing the features, we train an SVR model based on the dependency triples in the source language, and then apply the model to rank the triples in the target language. To avoid noisy annotations, we consider only those triples that occur at least twice in each corpus.

We further experiment with a simple ensemble method to combine source languages, in order to smooth over annotation and linguistic differences between languages: we combine the trained rerankers from multiple source languages by calculating the average of the predicted scores from each language.

[4] MWEs components are usually seen in a fixed order and with fixed gap size. We use mean to aggregate the pattern fixedness scores in order to capture any lexical instances which are not used in a fixed order. However, we use the median for the other measures to suppress the impact of outliers. Our preliminary results also confirm that this is the best way to aggregate the scores.

Score	Pattern	Example
0.327	$\langle$NOUN, ccomp, ADJ$\rangle$	*sure place*
0.306	$\langle$X, compound, PROPN$\rangle$	*Indo Lanka*
0.301	$\langle$NOUN, appos, SYM$\rangle$	*$ value*
0.298	$\langle$ADJ, nmod:npmod, ADV$\rangle$	*little more*
0.295	$\langle$NOUN, punct, NUM$\rangle$	*5 ”*
0.285	$\langle$SYM, punct, SYM$\rangle$	— —
0.283	$\langle$AUX, advcl, ADV$\rangle$	*as can*
0.277	$\langle$NOUN, mwe, SCONJ$\rangle$	*in case*
0.270	$\langle$SCONJ, mwe, ADP$\rangle$	*due to*
0.268	$\langle$NOUN, mwe, ADP$\rangle$	*in order*

Table 1: The top predicted MWE patterns in English, by combing all other languages.

6 Results

We report on two experiments. First, we train a model using features extracted from the gold-standard treebank in a given source language, and apply it to features extracted from the gold-standard treebank in a target language. We investigate how well our model is able to find the annotated MWE triples when gold-standard dependency relations are provided. This experiment also shows how our model can be used to find new MWE patterns in existing annotated treebanks (missing certain MWE types). Second, we investigate how our model performs in the more realistic scenario of no annotated treebank being available in the target language.

6.1 Experiment I: Learning given the gold standard treebank

In our first experiment, we assume access to gold standard annotations of POS tags and relation edges in both source and target languages, to determine the tractability of the task, assuming perfect parses.

Since the output of our model is a score in the range $[0, 1]$, we evaluate based on the area under the curve (AUC) from a ROC curve. Figure 2 shows the ROC curve for predicting MWEs when English and Norwegian are the source languages. English is among our 5 selected languages — i.e., one of the languages with the highest number of multiword expression patterns — while for Norwegian, the `mwe` relation is not used at all and only `compound:prt` is annotated. According to these results, the average AUC for predicting MWE patterns is 0.63 when English is the source language (averaged across all target languages, excluding English), while it is 0.50 when Norwegian is the source language. This shows that a source language with less annotated patterns makes for a weaker model. The average scores when our 5 selected languages are used as the source language are remarkably similar: English = 0.63, Persian = 0.64, Croatian = 0.64, Indonesian = 0.61 and Swedish = 0.65.

To investigate further, Figure 3 shows how adding more source languages affects the results for MWE pattern extraction over our 5 selected languages. According to these results, using more than one language can increase the AUC, however, using more than 3 languages does not improve the average AUC greatly.

Finally, we show the top-predicted MWE patterns in English in Table 1. We observe errors such as $\langle$NOUN, punct, NUM$\rangle$ and $\langle$SYM, punct, SYM$\rangle$, because of their idiosyncratic properties across token instances. However, our model also predicts that $\langle$NOUN, ccomp, ADJ$\rangle$ is an MWE.

6.2 Experiment II: Learning without gold standard dependency relations

In our second experiment, we evaluate under the more realistic task setting of there being no gold standard treebank in the target language. Instead, we use the cross-lingual parser proposed by Duong et al. (2015) to parse the corpus in the target language (see Section 2). Note that we still use gold-standard POS tags, but this isn't entirely unrealistic given the relative maturity of methods for inducing universal POS taggers (Das and Petrov, 2011; Täckström et al., 2013; Duong et al., 2014).

Obviously, due to the fact that the parser has no access to dependency annotations in the target language, the parser output will be noisy. However, this emulates a true surprise language setup, where we

Language	De	Sv	Da	It	Es	Fr	Hr	Bg	Hu	Fa	Ga	Eu	Id	He
Baseline	**0.816**	0.511	0.481	0.637	0.546	0.656	0.487	0.554	0.406	0.467	0.586	0.497	0.473	0.472
Reranker	0.736	**0.514**	0.428	0.631	**0.610**	**0.684**	0.461	**0.645**	0.470	**0.549**	0.673	0.578	0.513	**0.622**
PMI	0.804	0.512	**0.567**	**0.803**	0.494	0.634	**0.575**	0.471	**0.588**	0.521	**0.687**	**0.595**	**0.756**	0.476
Baseline + gold	0.797	0.750	0.846	0.885	0.879	0.799	0.696	0.917	0.457	0.919	0.799	0.891	0.931	0.788

Table 2: AUC scores when English is used as the source language to transfer dependency links and to train our reranker model. In "Baseline + gold", the trained model is applied to the gold-standard annotation of the target language rather than the parsed corpus.

have no prior knowledge of MWEs or dependency structure in the target language.

In order to evaluate our proposed method and compare it with the gold standard treebank, we change the evaluation method slightly in order to better reflect the expected inconsistencies in the parser output. In terms of gold-standard labeling, we exhaustively consider every edge between all pairs of tokens in each sentence, and consider an edge to be a positive instance if there is an MWE dependency between its token pairs in the gold-standard treebank, and a negative instance otherwise. To evaluate the parser output, which is the baseline in this experiment, we use the generated dependency edges and labels, and evaluate this against the exhaustively-generated gold-standard. To evaluate our system's performance, we use the dependency edges given by the parser and aggregate the reranker's predicted scores at the level of the delexicalised dependency triples, as per the first experiment. Unlike the first experiment, we evaluate using ROC AUC over the token pairs instead of $\langle \mathrm{pos}_h, \mathsf{rel}, \mathrm{pos}_d \rangle$ triples.

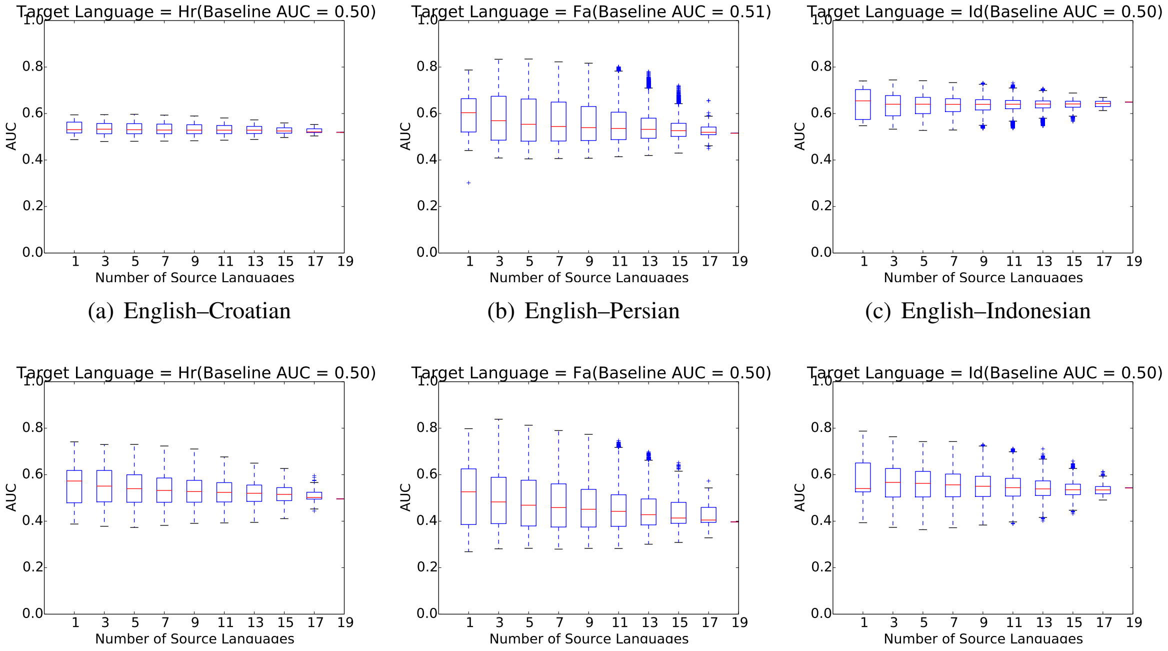
Table 2 shows the AUC scores when English is used as the source language to parse the target language, and English is also used to train our reranking model. Our proposed model ("Reranker") produces above-baseline results for all target languages except German, Danish, Italian and Croatian. We observe a very high percentage (63%) of compounds being predicted as noun–noun compounds in German, which is a large part of the strong results for that language. In order to compare with a collocation extraction methods, we contrast this with a ranking based on the average PMI score for each dependency relation ("PMI"). The results show that for half of the languages simple PMI scores can lead to higher AUC scores, while for the other half, the reranker model (which incorporates PMI scores but is trained on another language), performs better.

The final row in Table 2 is the result of providing the baseline method with gold standard dependency relations (with unknown label, to avoid trivialising the task) and applying the reranker to the gold-standard tuples. Since one source of noise (i.e. the induced parser) is removed in this baseline, we observe much higher scores than the other two approaches, except for Hungarian. For Hungarian, 90% of the annotated MWEs are NUM–NUM compounds, which is the reason that our second baseline performs worse for Hungarian compared to other languages. This result suggests that, unsurprisingly perhaps, the major cause of error in our method is the dependency parser.

Similar to the previous experiment, we also experimented with an ensemble of rerankers. We use English and Swedish as the source language to parse Persian, Croatian and Indonesian. Figure 4 shows how increasing the number of source languages and combining the trained models affects the AUC scores. According to Figure 4, our proposed method on average beats the baseline, when using only one language to train the reranker. However, unlike the previous experiment, combining multiple source languages does not improve the reranker. Additionally, comparing English with Swedish, we observe that the source language used to induce the dependency parser plays an important role.

7 Error Analysis

Finally, we perform error analysis to better understand the performance of the proposed method, focusing on two languages: Persian and Croatian, with 322 and 225 patterns to rank, respectively. We selected these two languages primarily because of the diversity of the MWE annotations in the treebanks (Section 3), and we had access to expert native-speaker annotators. The most frequent annotated patterns in the original treebanks are shown as "Gold standard" in Table 3 (the relation between pos_h and pos_d are either `mwe` or `compound` or both). The dependency parser and reranker are trained on English and



(a) English–Croatian

(b) English–Persian

(c) English–Indonesian

(d) Swedish–Croatian

(e) Swedish–Persian

(f) Swedish–Indonesian

Figure 4: Combining source languages given the noisy dependency relations. In (a)–(c), the English treebank is used as the source language for the cross-lingual parser, and in (d)–(f) Swedish is used.

Swedish as the source languages, individually. The top-10 most frequent patterns in the first quartile of the output of the reranker are shown in Table 3. The patterns which match with the gold standard pattern are marked with "†".

When English is the source language, noun compounds are correctly selected as a very frequent pattern in Persian. The pattern of ⟨ADJ, amod, NOUN⟩ is selected as the second most common pattern in Persian, for which almost all token instances are also true MWEs, such as *Islamic republic*, *Islamic revolution*, *right wing* and *fundamental law*.[5] Instances of ⟨NOUN, nmod, NOUN⟩ are more institutionalised, such as *seminar presentation* and *Iftar time*. Persian is rich with LVCs, which shows up in the first column as ⟨NOUN, nsubj, VERB⟩, i.e. misanalysed as verb–subject rather than verb–object pairs, but containing predominantly LVCs. In fact, among the top-20 most frequent token instances of this pattern, 17 are LVCs. As we work our way down the list of dependency triples in Table 3, there are fewer and fewer actual MWE token instances associated with the pattern. For example, the number of MWE instances associated with ⟨ADV, advmod, NOUN⟩ is less than non-MWEs (MWE examples are *before Christ*, and *before revolution*). The primary sources of error were parser errors or the triple being a fragment of a larger MWE. Using Swedish as the source language, we observed a similar trend.

For Croatian, almost all of the tokens associated with the top-2 patterns for both English and Swedish are MWE instances, with the tokens associated with ⟨NOUN, compound, NOUN⟩ based on English corresponding very closely with ⟨NOUN, nmod, NOUN⟩ based on Swedish. As with Persian, as we go down the list, the patterns become more noisy and the MWE tokens sparser, with the exception of ⟨NOUN, compound/nmod, PROPN⟩, for which almost all instances are MWEs (e.g. *president Erdogan*) or part of a larger MWE. Also, the instances of ⟨PROPN, compound, PROPN⟩ are all named entities. None of the token instances associated with ⟨NOUN, nsubj, VERB⟩ and ⟨AUX, aux, VERB⟩ were MWEs.

[5] Note that our model predicts the MWE patterns rather than MWE instances. Therefore, whether an individual MWE is also an MWE in the target language or not, does not affect the final results.

	Source = English	Source = Swedish	Gold standard
Persian	⟨NOUN, compound, NOUN⟩ †	⟨ADP, case, NOUN⟩	⟨NOUN, ∗, VERB⟩
	⟨ADJ, amod, NOUN⟩ †	⟨NOUN, nsubj, VERB⟩ †	⟨ADP, ∗, ADP⟩
	⟨NOUN, nmod, NOUN⟩ †	⟨NOUN, nmod:poss, NOUN⟩ †	⟨ADJ, ∗, VERB⟩
	⟨NOUN, nsubj, VERB⟩ †	⟨NOUN, nmod, VERB⟩ †	⟨NOUN, ∗, NOUN⟩
	⟨NOUN, nmod, ADJ⟩	⟨VERB, acl:relcl, NOUN⟩	⟨NUM, ∗, NOUN⟩
	⟨DET, det, NOUN⟩	⟨ADJ, amod, NOUN⟩ †	⟨CONJ, ∗, PRON⟩
	⟨ADV, advmod, NOUN⟩	⟨CONJ, cc, NOUN⟩ †	⟨CONJ, ∗, CONJ⟩
	⟨NOUN, conj, VERB⟩ †	⟨ADJ, nsubj, VERB⟩ †	⟨ADJ, ∗, NOUN⟩
	⟨NOUN, nmod, VERB⟩ †	⟨DET, det, NOUN⟩	⟨NOUN, ∗, ADP⟩
	⟨NOUN, conj, SCONJ⟩	⟨NUM, nummod, NOUN⟩ †	⟨CONJ, ∗, NOUN⟩
Croatian	⟨ADJ, amod, NOUN⟩ †	⟨ADJ, amod, NOUN⟩ †	⟨PRON, ∗, VERB⟩
	⟨NOUN, compound, NOUN⟩ †	⟨NOUN, nmod, NOUN⟩ †	⟨ADJ, ∗, NOUN⟩
	⟨ADP, case, NOUN⟩ †	⟨ADP, case, NOUN⟩ †	⟨NOUN, ∗, NOUN⟩
	⟨NOUN, nmod, NOUN⟩ †	⟨NOUN, dobj, VERB⟩	⟨PROPN, ∗, PROPN⟩
	⟨NOUN, dobj, VERB⟩	⟨NOUN, nmod, PROPN⟩	⟨PROPN, ∗, NOUN⟩
	⟨NOUN, compound, PROPN⟩	⟨NOUN, nmod:poss, NOUN⟩ †	⟨NUM, ∗, NOUN⟩
	⟨NOUN, nmod, VERB⟩	⟨NOUN, nmod, VERB⟩	⟨ADP, ∗, NOUN⟩
	⟨PROPN, compound, PROPN⟩ †	⟨NOUN, nsubj, VERB⟩	⟨X, ∗, X⟩
	⟨NOUN, nsubj, VERB⟩	⟨AUX, aux, VERB⟩	⟨PRON, ∗, ADP⟩
	⟨NOUN, conj, NOUN⟩ †	⟨PRON, nsubj, VERB⟩ †	⟨PRON, ∗, SCONJ⟩

Table 3: Top-ranking Persian and Croatian MWE patterns extracted using English and Swedish as the source language. Those patterns which match the top-ranking gold standard patterns are shown with "†".

8 Conclusion

In this paper, we proposed a method for automatically determining the MWE composition of a novel language, based on delexicalised universal dependency patterns of the form $\langle \text{pos}_h, \mathsf{rel}, \text{pos}_d \rangle$. The method is based on determination of MWEs in a source language from a dependency treebank, and training of a model over delexicalised dependency patterns for that language. This is then applied to a target language to rerank patterns, in terms of MWEhood. In our initial experiments, we used gold-standard dependency information for a treebank for the target language, and found the method to be highly successful at ranking dependency patterns. This both validated the method, as well as suggesting the potential for the use of the method in cross-checking the consistency of the UD treebanks. We then applied our method under the more realistic setting of having no gold-standard dependency data for the target language, but instead the output of a dependency parser induced for the target language based only on a POS-tagged monolingual corpus in the target language (and gold-standard data in the source language). We found the method to produce above-baseline results for the majority of languages tested, and that for the false positives associated with higher token frequencies, many of the associated tokens were actually true instances of MWEs (with the wrong dependency relation).

Acknowledgements

We wish to thank Long Duong for help with the transfer-based dependency parsing, Jan Snajder for his kind assistance with the Croatian annotation, and Dan Flickinger, Lars Hellan, Ned Letcher and João Silva for valuable advice in the early stages of development of this work. We would also like to thank the anonymous reviewers for their insightful comments and valuable suggestions. NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References

Josep Alba-Salas. 2002. *Light Verb Constructions in Romance. A Syntactic Analysis*. Ph.D. thesis, Cornell University.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 98–104, Taipei, Taiwan.

Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. 2004. In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier, editor, *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*. Kluwer Academic, Dordrecht, Netherlands.

Timothy Baldwin. 2005. The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

James Breen and Timothy Baldwin. 2009. Corpus-based extraction of Japanese compound verbs. In *Proceedings of the Australasian Language Technology Workshop 2009 (ALTW 2009)*, pages 35–43, Sydney, Australia.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 753–761, Dublin, Ireland.

Miriam Butt. 2003. The light verb jungle. In *Proceedings of the Workshop on Multi-Verb Constructions*, pages 1–49, Trondheim, Norway.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 204–212, Jeju Island, Korea.

Ann Copestake. 2003. Compounds revisited. In *Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon*, Geneva, Switzerland.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, USA.

Mona T. Diab and Pravin Bhutada. 2009. Verb noun construction MWE token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 17–22, Singapore.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 886–897, Doha, Qatar.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the 19th Conference on Natural Language Learning (CoNLL-2015)*, pages 113–122, Beijing, China.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):450–466.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 138–147, Atlanta, USA.

Fei Huang, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, Sapporo, Japan.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Philadelphia, USA.

Gholam Hossein Karimi-Doostan. 1997. *Light Verb Construction in Persian*. Ph.D. thesis, University of Essex.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, pages 39–46, Toulouse, France.

Mirella Lapata and Alex Lascarides. 2003. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics (EACL-2003)*, pages 235–242, Budapest, Hungary.

Linlin Li and Caroline Sporleder. 2010. Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Posters Volume*, pages 683–691, Beijing, China.

Diana Maynard, Valentin Tablan, Kalina Bontcheva, and Hamish Cunningham. 2003. Rapid customization of an information extraction system for a surprise language. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):295–300.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the Fifth Workshop on Very Large Corpora*. EMNLP.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*, Scottsdale, USA.

Begona Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-wordexpressions in a multilingual context*, pages 33–40.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 629–637, Jeju Island, Korea.

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2077–2092, Mumbai, India.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Douglas W Oard. 2003. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):79–84.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge.

Pavel Pecina. 2008. *Lexical Association Measures: Collocation Extraction*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey.

Carlos Ramisch. 2012. A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, pages 61–66, Jeju Island, Korea.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing Computational Linguistics (CICLing-2002)*, pages 189–206, Mexico City, Mexico.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association of Computational Linguistics*, 2:193–206.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 255–265, Atlanta, USA.

Sabine Schulte im Walde. 2004. Identification, quantitative description, and preliminary distributional analysis of German particle verbs. In *Proceedings of the COLING Workshop on Enhancing and Using Electronic Dictionaries*, pages 85–88, Geneva, Switzerland.

Ralf Steinberger and Bruno Pouliquen. 2007. Cross-lingual named entity recognition. *Lingvisticæ Investigationes*, 30(1):135–162.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 1061–1071, Atlanta, USA.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation a feasibility study on shallow processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):497–512.

Leonoor van der Beek. 2005. The extraction of determinerless PPs. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 190–199, Colchester, UK.

Sriram Venkatapathy and Aravind Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 53–60, Sydney, Australia.