Machine Learning of Syntactic Attachment from Morphosyntactic and Semantic Co-occurrence Statistics

Adam Slaski and **Szymon Acedański** and **Adam Przepiórkowski**
University of Warsaw
and
Institute of Computer Science
Polish Academy of Sciences

Abstract

The paper presents a novel approach to extracting dependency information in morphologically rich languages using co-occurrence statistics based not only on lexical forms (as in previously described collocation-based methods), but also on morphosyntactic and wordnet-derived semantic properties of words. Statistics generated from a corpus annotated only at the morphosyntactic level are used as features in a Machine Learning classifier which is able to detect which heads of groups found by a shallow parser are likely to be connected by an edge in the complete parse tree. The approach reaches the precision of 89% and the recall of 65%, with an extra 6% recall, if only words present in the wordnet are considered.

1 Introduction

The practical issue handled in this paper is how to connect syntactic groups found by a shallow parser into a possibly complete syntactic tree, i.e., how to solve the attachment problem. To give a well-known example from English, the task is to decide whether in *I shot an elephant in my pajamas*[1], the group *in my pajamas* should be attached to *an elephant* or to *shot* (or perhaps to *I*).

The standard approach to this problem relies on finding collocation strengths between syntactic objects, usually between lexical items which are heads of these objects, and resolve attachment ambiguities on the basis of such collocation information. The current work extends this approach in two main ways. First, we consider a very broad range of features: not only lexical, but also lexico-semantic, lexico-grammatical, and grammatical. Second, and more importantly, we train classifiers based not on these features directly, but rather on various association measures calculated for each of the considered features. This way the classifier selects which types of features are important and which association measures are most informative for any feature type.

The proposed method is evaluated on Polish, a language with rich inflection (and relatively free word order), which exacerbates the usual data sparseness problem in NLP.

In this work we assume that input texts are already part-of-speech tagged and chunked, the latter process resulting in the recognition of basic syntactic groups. A syntactic group may, e.g., consist of a verb with surrounding adverbs and particles or a noun with its premodifiers. We assume that all groups have a syntactic head and a semantic head. In verbal and nominal groups both heads are the same word, but in prepositional and numeral groups they usually differ: the preposition and the numeral are syntactic heads of the respective constituents, while the semantic head is the head noun within the nominal group contained in these constituents.

To simplify some of the descriptions below, by *syntactic object* we will understand either a shallow group or a word. We will also uniformly talk about syntactic and semantic heads of all syntactic objects; in case of words, the word itself is its own syntactic and semantic head. In effect, any syntactic object may be represented by a pair of words (the two

[1] http://www.youtube.com/watch?v=NfN_gcjGoJo

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 42–47,
Jeju, Republic of Korea, 12 July 2012.

heads), and each word is characterised by its base form and its morphosyntactic tag.

2 Algorithm

The standard method of solving the PP-attachment problem is based on collocation extraction (cf., e.g., (Hindle and Rooth, 1993)) and consists of three main steps: first a training corpus is scanned and frequencies of co-occurrences of pairs of words (or more general: syntactic objects) are gathered; then the collected data are normalised to obtain, for each pair, the strength of their connection; finally, information about such collocation strengths is employed to solve PP-attachment in new texts. An instance of the PP-attachment problem is the choice between two possible edges in a parse tree: (n_1, pp) and (n_2, pp), where pp is the prepositional phrase, and n_1 and n_2 are nodes in the tree (possible attachment sites). This is solved by choosing the edge with the node that has a stronger connection to the pp.

On this approach, collocations (defined as a relation between lexemes that co-occur more often than would be expected by chance) are detected by taking pairs of syntactic objects and only considering the lemmata of their semantic heads. The natural question is whether this could be generalised to other properties of syntactic objects. In the following, the term *feature* will refer to any properties of linguistic objects taken into consideration in the process of finding collocation strengths between pairs of objects.

2.1 Lexical and Morphosyntactic Features

To start with an example of a generalised collocation, let us consider morphosyntactic valence. In order to extract valence links between two objects, we should consider the lemma of one object (potential predicate) and the morphosyntactic tag, including the value of case, etc., of the other (potential argument). This differs from standard (lexical) collocation, where the same properties of both objects are considered, namely, their lemmata.

Formally, we define a feature f to be a pair of functions $l_f \colon \mathsf{so} \to L_f$ and $r_f \colon \mathsf{so} \to R_f$, where so stands for the set of syntactic objects and L_f, R_f are the investigated properties. For example, to learn dependencies between verbs and case values of their objects, we can take $l_f(\mathsf{w}) = \mathsf{base}(\mathsf{semhead}(\mathsf{w}))$ (the lemma of the semantic head of w) and $r_f(\mathsf{w}) = \mathsf{case}(\mathsf{synhead}(\mathsf{w}))$ (the case value of the syntactic head of w). On the other hand, in order to obtain the usual collocations, it is sufficient to take both functions as mapping a syntactic object to a base form of its semantic head.

What features should be considered in the task of finding dependencies between syntactic objects? The two features mentioned above, aimed at finding lexical collocations and valence relations, are obviously useful. However, in a morphologically rich language, like Polish, taking the full morphosyntactic tag as the value of a feature function leads to the data sparsity problem. Clearly, the most important valence information a tag may contribute is part of speech and grammatical case. Hence, we define the second function in the "valence" feature more precisely to be the base form and grammatical case (if any), if the syntactic object is a preposition, or part of speech and grammatical case (if any), otherwise. For example, consider the sentence *Who cares for the carers?* and assume that it has already been split into basic syntactic objects in the following way: [*Who*] [*cares*] [*for the carers*] [*?*]. The syntactic head of the third object is *for* and the lemma of the semantic head is CARER. So, the valence feature for the pair *care* and *for the carers* (both defined below via their syntactic and semantic heads) will give:

$$l_{val}(\langle \textsc{care}{:}\mathsf{verb}, \mathsf{3s}; \textsc{care}{:}\mathsf{verb}, \mathsf{3s}\rangle) = \textsc{care}$$
$$r_{val}(\langle \textsc{for}{:}\mathsf{prep}, \mathsf{obj}; \textsc{carer}{:}\mathsf{noun}, \mathsf{pl}\rangle) = \langle \textsc{for}, \mathsf{obj}\rangle,$$

where $\mathsf{3s}$ stands for the "3rd person singular" property of verbs and obj stands for the objective case in English.

Additionally, 7 morphosyntactic features are defined by projecting both syntactic objects onto any (but the same of the two objects) combination of grammatical case, gender and number. For example one of those features is defined in the following way:

$$l_f(\mathsf{w}) = r_f(\mathsf{w}) = \\ = \langle \mathsf{case}(\mathsf{synhead}(\mathsf{w})), \mathsf{gender}(\mathsf{synhead}(\mathsf{w}))\rangle.$$

Another feature relates the two objects' syntactic heads, by looking at the part of speech of the first one and the case of the other one. The final feature records information about syntactic (number, gender, case) agreement between the objects.

2.2 Lexico-Semantic Features

Obviously, the semantics of syntactic objects is important in deciding which two objects are directly connected in a syntactic tree. To this end, we utilise a wordnet.

Ideally, we would like to represent a syntactic object via its semantic class. In wordnets, semantic classes are approximated by synsets (synonym sets) which are ordered by the hyperonymy relation. We could represent a syntactic object by its directly corresponding synset, but in terms of generalisation this would hardly be an improvement over representing such an object by its semantic head. In most cases we need to represent the object by a hypernym of its synset. But how far up should we go along the hypernymy path to find a synset of the right granularity? This is a difficult problem, so we leave it to the classifier. Instead, lexico-semantic features are defined in such a way that, for a given lexeme, all its hypernyms are counted as observations.

After some experimentation, three features based on this idea are defined:

1. $l_f(\mathsf{w}) = \mathsf{base}(\mathsf{semhead}(\mathsf{w}))$
   $r_f(\mathsf{w}) = \mathsf{sset}(\mathsf{w})$
   (for all $\mathsf{sset}(\mathsf{w}) \in \mathsf{hypernyms}(\mathsf{w})$),
2. $l_f(\mathsf{w}) = \mathsf{base}(\mathsf{semhead}(\mathsf{w}))$
   $r_f(\mathsf{w}) = \langle \mathsf{sset}(\mathsf{w}), \mathsf{case}(\mathsf{synhead}(\mathsf{w}))\rangle$
   (for all $\mathsf{sset}(\mathsf{w}) \in \mathsf{hypernyms}(\mathsf{w})$),
3. $l_f(\mathsf{w}) = \mathsf{sset}(\mathsf{w})$
   $r_f(\mathsf{w}) = \mathsf{sset}(\mathsf{w})$

In the last feature, where both objects are represented by synsets, only those minimally general hypernyms of the two objects are considered that cooccur in the training corpus more than T (threshold) times. In the experiments described below, performed on a 1-million-word training corpus, the threshold was set to 30.

2.3 Association Measures

For any two syntactic objects in the same sentence the strength of association is computed between them using each of the 14 features (standard collocations, 10 morphosyntactic features, 3 lexico-semantic features) defined above. In fact, we use not 1 but 6 association measures most suitable for language analysis according to (Seretan, 2011): log likehood ratio, chi-squared, t-score, z-score, pointwise mutual information and raw frequency. The last choice may seem disputable, but as was shown in (Krenn and Evert, 2001) (and reported in various works on valence acquisition), in some cases raw frequency behaves better than more sophisticated measures.

We are well aware that some of the employed measures require the distribution of frequencies to meet certain conditions that are not necessarily fulfilled in the present case. However, as explained in the following subsection, the decision which measures should ultimately be taken into account is left to a classifier.

2.4 Classifiers

Let us first note that no treebank is needed for computing the features and measures presented in the previous section. These measures represent co-occurrence strengths of syntactic objects based on different grouping strategies (by lemma, by part of speech, by case, gender, number, by wordnet synsets, etc.). Any large, morphosyntactically annotated (and perhaps chunked) corpus is suitable for computing such features. A treebank is only needed to train a classifier which uses such measures as input signals.[2]

In order to apply Machine Learning classifiers, one must formally define what counts as an instance of the classification problem. In the current case, for each pair of syntactic objects in a sentence, a single instance is generated with the following signals:

- absolute distance (in terms of the number of sytnactic objects in between),
- ordering (the sign of the distance),
- 6 measures (see § 2.3) of lexical collocation,
- 10 × 6 = 60 values of morphosyntactic co-occurrence measures,
- 3 × 6 = 18 values of lexico-semantic (wordnet-based) co-occurrence measures,
- a single binary signal based on 14 high-precision low-recall handwritten syntactic de-

[2] We use the term *signal* instead of the more usual *feature* in order to avoid confusion with features defined in § 2.1 and in § 2.2.

cision rules which define common grammatical patterns like verb-subject agreement, genitive construction, etc.; the rules look only at the morphosyntactic tags of the heads of syntactic objects,

- the classification target from the treebank: a binary signal describing whether the given pair of syntactic objects form an edge in the parse tree.

The last signal is used for training the classifier and then for evaluation. Note that lexical forms of the compared syntactic objects or their heads are not passed to the classifier, so the size of the training treebank can be kept relatively small.

An inherent problem that needs to be addressed is the imbalance between the sizes of two classification categories. Of course, most of the pairs of the syntactic objects do not form an edge in the parse tree, so a relatively high classification accuracy may be achieved by the trivial classifier which finds no edges at all. We experimented with various well-known classifiers, such as decision trees, Support Vector Machines and clustering algorithms, and also tried subsampling[3] of the imbalanced data. Finally, satisfactory results were achieved by employing a Balanced Random Forest classifier.

Random Forest (Breiman, 2001) is a set of unpruned C4.5 (Quinlan, 1993) decision trees. When building a single tree in the set, only a random subset of all attributes is considered at each node and the best is selected for splitting the data set. Balanced Random Forest (BRF, (Chen et al., 2004)) is a modified version of the Random Forest. A single tree of BRF is built by first randomly subsampling the more frequent instances in the training set to match the number of less frequent ones and then creating a decision tree from this reduced data set.

3 Experiments and Evaluation

The approach presented above has been evaluated on Polish.

First, a manually annotated 1-million-word subcorpus of the National Corpus of Polish (Przepiórkowski et al., 2010), specifically, its morphosyntactic and shallow syntactic annotation, was used to compute the co-occurrence statistics. The wordnet used for lexico-semantic measures was *Słowosieć* (Piasecki et al., 2009; Maziarz et al., 2012), the largest Polish wordnet.

Then a random subset of sentences from this corpus was shallow-parsed by Spejd (Buczyński and Przepiórkowski, 2009) and given to linguists, who added annotation for the dependency links between syntactic objects. Each sentence was processed by two linguists, and in case of any discrepancy, the sentence was simply rejected. The final corpus contains 963 sentences comprising over 8000 tokens.

From this data we obtained over 23 500 classification problem instances. Then we performed the classification using a BRF classifier written for Weka (Witten and Frank, 2005) as part of the research work on definition extraction with BRFs (Kobyliński and Przepiórkowski, 2008). The results were 10-fold cross-validated. A similar experiment was performed taking into account only those instances which describe syntactic objects with semantic heads present in the wordnet. The results were measured in terms of precision and recall over edges in the syntactic tree: what percentage of found edges are correct (precision) and what percentage of correct edges were found by the algorithm (recall). The obtained measures are presented in Table 1.

Expected		
YES	NO	Classified
2674	319	YES
1781	21250	NO
	Precision:	0.89
	Recall:	0.60
	F-measure:	0.72

Expected		
YES	NO	Classified
1933	241	YES
1008	13041	NO
	Precision:	0.89
	Recall:	0.66
	F-measure:	0.76

Table 1: Confusion matrix (# of instances) and measures for the full data set and for data present in wordnet.

[3] Removing enough negative instances in the training set to balance the numbers of instances representing both classes.

We also looked at the actual decision trees that were generated during the training. We note that the signal most frequently observed near the tops of decision trees was the one from handwritten rules. The second one was the distance. By looking at the trees, we could not see any clear preferences for other types of signals. This suggests that both morphosyntactic and lexico-semantic signals contribute to the accuracy of the classification.

Based on this inspection of decision trees, we performed another experiment to learn how much improvement we get from generalised collocation signals. We evaluated – on the same data – a not so trivial baseline algorithm which, for each syntactic object, creates an edge to its nearest neighbour accepted by the handwritten rules, if any. Note that this baseline builds on the fact that a node in a parse tree has at most one parent, whereas the algorithm described above does not encode this property, yet; clearly, there is still some room for improvement. The baseline reaches 0.78 precision and 0.47 recall (F-measure is 0.59). Therefore, the improvement from co-occurrence signals over this strong baseline is 0.13, which is rather high. Also, given the high precision, our algorithm may be suitable for using in a cascade of classifiers.

4 Related Work

There is a plethora of relevant work on resolving PP-attachment ambiguities in particular and finding dependency links in general, and we cannot hope to do it sufficient justice here.

One line of work, exemplified by the early influential paper (Hindle and Rooth, 1993), posits the problem of PP-attachment as the problem of choosing between a verb v and a noun n_1 when attaching a prepositional phrase defined by the syntactic head p and the semantic head n_2. Early work, including (Hindle and Rooth, 1993), concentrated on lexical associations, later also using wordnet information, e.g., (Clark and Weir, 2000), in a way similar to that described above. Let us note that this scenario was criticised as unrealistic by (Atterer and Schütze, 2007), who argue that "PP attachment should not be evaluated in isolation, but instead as an integral component of a parsing system, without using information from the gold-standard oracle", as in the approach proposed here.

Another rich thread of relevant research is concerned with valence acquisition, where shallow parsing and association measures based on morphosyntactic features are often used at the stage of collecting evidence, (Manning, 1993; Korhonen, 2002), also in work on Polish, (Przepiórkowski, 2009). However, the aim in this task is the construction of a valence dictionary, rather than disambiguation of attachment possibilities in a corpus.

A task more related to the current one is presented in (Van Asch and Daelemans, 2009), where a PP-attacher operates on top of a shallow parser. However, this memory-based module is fully trained on a treebank (Penn Treebank, in this case) and is concerned only with finding anchors for PPs, rather than with linking any dependents to their heads.

Finally, much work has been devoted during the last decade to probabilistic dependency parsing (see (Kübler et al., 2009) for a good overview). Classifiers deciding whether – at any stage of dependency parsing – to perform *shift* or *reduce* typically rely on lexical and morphosyntactic, but not lexico-semantic information (Nivre, 2006). Again, such classifiers are fully trained on a treebank (converted to parser configurations).

5 Conclusion

Treebanks are very expensive, morphosyntactically annotated corpora are relatively cheap. The main contribution of the current paper is a novel approach to factoring out syntactic training in the process of learning of syntactic attachment. All the fine-grained lexical training data were collected from a relatively large morphosyntactically annotated and chunked corpus, and only less than 100 signals (although many of them continuous) were used for training the final classifier on a treebank. The advantage of this approach is that reasonable results can be achieved on the basis of tiny treebanks (here, less than 1000 sentences).

We are not aware of work fully analogous to ours, either for Polish or for other languages, so we cannot fully compare our results to the state of the art. The comparison with a strong baseline algorithm which uses handwritten rules shows a significant improvement – over 0.13 in terms of F-measure.

Acknowledgments

This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund.

References

Michaela Atterer and Hinrich Schütze. 2007. Prepositional phrase attachment without oracles. *Computational Linguistics*, 33(4):469–476.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Aleksander Buczyński and Adam Przepiórkowski. 2009. Spejd: A shallow processing and morphological disambiguation tool. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human Language Technology: Challenges of the Information Society*, volume 5603 of *Lecture Notes in Artificial Intelligence*, pages 131–141. Springer-Verlag, Berlin.

Chao Chen, Andy Liaw, and Leo Breiman. 2004. Using random forest to learn imbalanced data. Technical Report 666, University of California, Berkeley.

Stephen Clark and David Weir. 2000. A class-based probabilistic approach to structural disambiguation. In *In Proceedings of the 18th International Conference on Computational Linguistics*, pages 194–200.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Łukasz Kobyliński and Adam Przepiórkowski. 2008. Definition extraction with balanced random forests. In Bengt Nordström and Aarne Ranta, editors, *Advances in Natural Language Processing: GoTAL 2008, Gothenburg, Sweden*, volume 5221 of *Lecture Notes in Artificial Intelligence*, pages 237–247, Berlin. Springer-Verlag.

Anna Korhonen. 2002. *Subcategorization Acquisition*. PhD Thesis, University of Cambridge.

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 Workshop on Collocations*, Toulouse, France.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, OH.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan.

Joakim Nivre. 2006. *Inductive Dependency Parsing*. Springer-Verlag, Berlin.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. 2010. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Adam Przepiórkowski. 2009. Towards the automatic acquisition of a valence dictionary for Polish. In Małgorzata Marciniak and Agnieszka Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *Lecture Notes in Computer Science*, pages 191–210. Springer-Verlag, Berlin.

John Ross Quinlan. 1993. *C4.5 Programs for Machine Learning*. Morgan Kaufmann.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer, Dordrecht.

Vincent Van Asch and Walter Daelemans. 2009. Prepositional phrase attachment in shallow parsing. In *Proceedings of the International Conference RANLP-2009*, pages 12–17, Borovets, Bulgaria, September.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA, 2nd edition.