From a Distance: Using Cross-lingual Word Alignments for Noun Compound Bracketing

Patrick Ziering
Institute for NLP
University of Stuttgart, Germany

Lonneke van der Plas
Institute of Linguistics
University of Malta, Malta

Patrick.Ziering@ims.uni-stuttgart.de
Lonneke.vanderPlas@um.edu.mt

Abstract

We present a cross-lingual method for determining NP structures. More specifically, we try to determine whether the semantics of tripartite noun compounds in context requires a left or right branching interpretation. The system exploits the difference in word position between languages as found in parallel corpora. We achieve a bracketing accuracy of 94.6%, significantly outperforming all systems in comparison and comparable to human performance. Our system generates large amounts of high-quality bracketed NPs in a multilingual context that can be used to train supervised learners.

1 Introduction

k-partite noun compounds, i.e., compositions of k bare common nouns that function as one unit, (kNCs), such as *air traffic control system*, usually have an implicit structure that reflects semantics. While a LEFT-branching *[world banana] market* is very unlikely, for *luxury cattle truck*, both structures make sense and context is necessary for disambiguation: *[luxury cattle] truck* is a truck for luxury cattle whereas *luxury [cattle truck]* is a luxury truck for (any) cattle. Therefore, a proper structural analysis is a crucial part of noun compound interpretation and of fundamental importance for many tasks in natural language processing such as machine translation. The correct French translation of *luxury cattle truck* depends on the internal structure. While [*luxury cattle*] *truck* is translated as *camion pour bétail de luxe*, the preferred translation for *luxury* [*cattle truck*] is *camion de luxe pour bétail*.

Previous work on noun compound bracketing has shown that supervised beats unsupervised. The latter approaches use N-gram statistics or lexical patterns (Lauer, 1995; Nakov and Hearst, 2005; Barrière and Ménard, 2014), web counts (Lapata and Keller, 2004) or semantic relations (Kim and Baldwin, 2013) and evaluate on carefully selected evaluation data from encyclopedia (Lauer, 1995; Barrière and Ménard, 2014) or from general newspaper text (Kim and Baldwin, 2013). Vadas and Curran (2007a,b) manually annotated the Penn Treebank and showed that they improve over unsupervised results by a large margin. Pitler et al. (2010) used the data from Vadas and Curran (2007a) for a parser applicable on base noun phrases (NPs) of any length including coordinations. Barker (1998) presents a bracketing method for k-partite NPs that reduces the task to three-word bracketings within a sliding window. One advantage of supervised approaches for this task is that kNCs are labeled in context so contextual features can be used in the learning framework. These are especially useful when dealing with ambiguous kNCs.

The need for annotated data is a drawback of supervised approaches. Manual annotations are costly and time-consuming. To circumvent this need for annotated data, previous work has used cross-lingual supervision based on parallel corpora. Bergsma et al. (2011) made use of small amounts of annotated data on the target side and complement this with bilingual features from unlabeled bitext in a co-trained classifier for coordination disambiguation in complex NPs. Previous work on using cross-lingual data for the analysis of multi-word expressions (MWEs) of different types include Busa and Johnston (1996); Girju (2007); Sinha (2009); Tsvetkov and Wintner (2010); Ziering et al. (2013).

Proceedings of the 11th International Conference on Computational Semantics, pages 82–87,
London, UK, April 15-17 2015.

Ziering and Van der Plas (2014) propose an approach that refrains from using any human annotation. They use the fact, that languages differ in their preference for open or closed compounding (i.e., multiword vs. one-word compounds), for inducing the English bracketing of 3NCs. English open 3NCs like *human rights abuses* can be translated to partially closed phrases as in German *Verletzungen der Menschenrechte*, (*abuses of human rights*), from which we can induce the LEFT-branching structure. Although this approach achieves a solid accuracy, a crucial limitation is coverage, because restricting to six paraphrasing patterns ignores many other predictive cases. Moreover, the system needs part of speech (PoS) tags and splitting information for determining 2NCs and is therefore rather language-dependent.

In this paper, we present a precise, high-coverage and knowledge-lean method for bracketing kNCs (for $k \geq 3$) occurring in parallel data. Our method uses the distances of words that are aligned to kNC components in parallel languages. For example, the 3NC *human rights violations* can be bracketed using the positions of aligned words in the Italian fragment ... *che le* ***violazioni*** *gravi e sistematiche dei* ***diritti umani*** .... The fact, that the alignment of the third noun, *violations* (**violazioni**), is separated from the rest, points us in the direction of LEFT-branching. Using less restricted forms of cross-lingual supervision, we achieve a much higher coverage than Ziering and Van der Plas (2014). Furthermore, our results are more accurate. In contrast to previous unsupervised methods, our system is applicable in both token- and type-based modes. Token-based bracketing is context-dependent and allows for a better treatment of structural ambiguity (as in *luxury cattle truck*). We generate large amounts of high-quality bracketed kNCs in a multilingual context that can be used to train supervised learners.

2 Aligned Word Distance Bracketing

The aligned word distance bracketing (AWDB) is inspired by Behaghel's First Law saying that elements which belong close together intellectually will also be placed close together (Behaghel, 1909).

1: $c_1, \ldots, c_n \Leftarrow \mathrm{N}_1, \ldots, \mathrm{N}_k$
2: $AW_i \Leftarrow$ set of content words c_i aligns to
3: **while** $|\{c_1, \ldots, c_n\}| > 1$ **do**
4: $\quad (c_m, c_{m+1}) \Leftarrow c$-pair with minimal AWD
5: $\quad$ merge c_m and c_{m+1} to $c_{[m,\, m+1]}$
6: $\quad AW_{[m,\, m+1]} = AW_m \cup AW_{m+1}$
7: **end while**

Figure 1: AWDB algorithm for kNCs

For each language l, we apply the AWDB algorithm on a kNC as shown in Figure 1: we start bottom-up with one constituent per noun. For each constituent c_i, we create a set of content words[1] c_i aligns to, AW_i. We merge the two consecutive constituents c_m and c_{m+1} with the smallest aligned word distance (AWD) based on the minimum distance from all words in AW_m to all words in AW_{m+1}:

$$\mathrm{AWD}(c_m, c_{m+1}) = \min_{x \in AW_m, y \in AW_{m+1}} |pos(x) - pos(y)|$$

where $pos(\alpha)$ is the position of a word α in a sentence. In the case of (c_m, c_{m+1}) being aligned to a common closed compound, $\mathrm{AWD}(c_m, c_{m+1})$ is zero. If the smallest AWD is not unique but the related constituents do not overlap (e.g., (c_1,c_2) and (c_3,c_4) aligning to two different closed compounds) we merge both constituent pairs in one iteration. If they overlap (e.g., (c_1,c_2) and (c_2,c_3) aligning to a common closed compound), no bracketing structure can be derived from the word positions in l. Similarly, if there is an empty set AW_e, i.e., there is no alignment from c_e to a content word in l, AWDB cannot bracket the kNC using the translation to l. If no structure can be derived from any aligned language, AWDB refuses to answer.

For example, the 4NC *air transport safety organization* is aligned to four words in the French fragment *Nous devons mettre en place cette* ***organisation***$_7$ *européenne chargée de la* ***sécurité***$_{12}$ *du* ***transport***$_{14}$ ***aérien***$_{15}$ *qui* ... (*We need to establish this European* ***organization*** *responsible for the* ***safety*** *of* ***air transport*** *that* ...). The aligned word sets are: $AW_1 = \{$aérien$\}$, $AW_2 = \{$transport$\}$, $AW_3 = \{$sécurité$\}$ and

[1] These are words tagged as noun, adjective or verb. They can be identified with corpus frequency to remain knowledge-lean.

$AW_4 = \{\text{organisation}\}$. c_1 and c_2 have the smallest AWD and thus are merged. In the next iteration, the smallest AWD is between $c_{[1,\ 2]}$ and c_3. As last step, we merge $c_{[[1,\ 2],\ 3]}$ and c_4. The resulting constituent corresponds to the 4NC structure [[[*air transport*] *safety*] *organization*].

To determine the final bracketing for a given kNC, we use the majority vote of all structures derived from all aligned languages. In the case of a tie, AWDB does not produce a final bracketing. Although this decision leads to lower coverage, it enables us to measure the pure impact of the cross-lingual word distance feature. For practical purposes, an additional back-off model is put in place. In order to mitigate word alignment problems and data sparseness, we additionaly bracket kNCs in a type-based fashion, i.e., we collect all kNC structures of a kNC type from various contexts.

3 Experiments

Tools and data. While AWDB is designed for bracketing NPs of any length, we first experiment with bracketing 3NCs, the largest class of 3^+NCs (93.8% on the basic dataset of Ziering and Van der Plas (2014)), for which bracketing is a binary classification (i.e., LEFT or RIGHT). For bracketing longer NCs we often have to make do with partial information from a language, instead of a full structure. In future work, we plan to investigate methods to combine these partial results. Moreover, in contrast to previous work (e.g., Vadas and Curran (2007b)), we take only common nouns as components into account rather than named entities. We consider the task of bracketing 3NCs composed of common nouns more ambitious, because named entities often form a single concept that is easy to spot, e.g., ***Apple II*** *owners*. Although AWDB can also process compounds including adjectives (e.g., ***active*** *inclusion policy* aligned to the Dutch *beleid voor actieve insluiting* (*policy for active inclusion*)), for a direct comparison with the system of Ziering and Van der Plas (2014), that analyses 3NCs, we restrict ourselves to noun sequences.

We use the Europarl[2] compound database[3] developed by Ziering and Van der Plas (2014). This database has been compiled from the OPUS[4] corpus (Tiedemann, 2012) and comprises ten languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese, Spanish and Swedish. We use the initial version (basic dataset), that contains English word sequences that conform PoS chunks and their alignments. We select English word sequences whose PoS pattern conforms three nouns.

Extraction errors are a problem, since many adjectives have been tagged as nouns and some 3NCs occur as incomplete fragments. For increasing the effectiveness of human annotation, we developed a high-confidence noun filter $P_{noun}(word) = P(noun \mid word)$. It is trained on the English Wikipedia[5] tagged by TreeTagger (Schmid, 1995). We inspect all 3NCs in the context of one token to the left and right, $w_0\{N_1N_2N_3\}w_4$. If $P_{noun}(N_i) < \theta$ or $P_{noun}(w_j) \geq \theta$, we remove the 3NC from our dataset. We inspected a subset of all 3NCs in the database and estimated the best filter quality to be with $\theta = 0.04$. This threshold discards ***increasing*** *land abandonment* but keeps ***human*** *rights abuse*. Our final dataset contains 14,941 tokens and 8824 types.

Systems in comparison. We compare AWDB with the bracketing approach of Ziering and Van der Plas (2014). For both systems, we use the majority vote across all nine aligned languages, in a token- and type-based version. We implemented an unsupervised method based on statistics on bi-grams extracted from the English part of the Europarl corpus.[6] As scorer, we use the *Chi squared* (χ^2) measure, which worked best in previous work (Nakov and Hearst, 2005). We consider both the *adjacency* (i.e., (N_1, N_2) vs. (N_2, N_3), (Marcus, 1980)) and the *dependency* (i.e., (N_1, N_2) vs. (N_1, N_3), (Lauer, 1995)) model. We created a back-off model for the bracketing system of Ziering and Van der Plas (2014) and for AWDB that falls back to using χ^2 if no bracketing structure can be derived (*system* $\rightarrow \chi^2$). Finally, we compare with a baseline, that always predicts the majority class: LEFT.

Human annotation. We observed that there is only a very small overlap between test sets of previous work on NP bracketing and the Europarl database. The test set used by Ziering and Van der Plas (2014) is very small and the labeling is less fine-grained. Thus, we decided to create our own test set.

[2] statmt.org/europarl

[3] ims.uni-stuttgart.de/data/NCDatabase.html

[4] opus.lingfil.uu.se

[5] en.wikipedia.org

[6] For a fair comparison, we leave systems that have access to external knowledge, such as web search engines, aside.

A trained independent annotator classified a set of 1100 tokens in context with one of the following labels: LEFT, RIGHT, EXTRACTION (for extraction errors that survived the high-confidence noun filter $P_{noun}(word)$), UNDECIDED (for 3NCs that cannot be disambiguated within the one-sentence context) and SEMANTIC INDETERMINATE (for 3NCs for which LEFT and RIGHT have the same meaning such as *book price fixing* (i.e., *price fixing for books* is equivalent to *fixing of the book price*)). We consider the full dataset to compare the coverage of the systems in comparison. For the accuracy figures, in order to keep annotation efforts small, we asked evaluators to annotate just those tokens that our system provides an answer to, because tokens that our system has no answer to will not be evaluated in the comparative evaluation on accuracy anyhow. Two additional trained independent annotators each classified one half of the dataset for checking inter-annotator agreement. For the classes LEFT/RIGHT (308 tokens), we achieve an agreement rate of 90.3% and $\kappa = 0.717$ (Cohen, 1960), which means good agreement (Landis and Koch, 1977). We use the LEFT/RIGHT consensus of the 3NC tokens as final test set (278 tokens).

Evaluation Measure. We measure accuracy (Acc $_\Omega$) for a set of 3NC tokens, Ω, as correct LEFT/RIGHT labels divided by all LEFT/RIGHT labels. Coverage is measured as LEFT/RIGHT labels divided by all 3NC tokens in the full dataset. We refer to the harmonic mean of Acc $_\Omega$ and Coverage as $harmonic(\Omega)$.

4 Results and Discussion

System	Coverage
AWDB$_{token}$ / AWDB$_{type}$	87.9% / 91.2%
AWDB$_{type} \rightarrow \chi^2$	100%
χ^2	100%
Zier.v.d.Plas14$_{token}$ / Zier.v.d.Plas14$_{type}$	29.9% / 48.1%
Zier.v.d.Plas14$_{type} \rightarrow \chi^2$	100%
LEFT baseline	100%

Table 1: Evaluation results on coverage

As it turned out that the *adjacency* model outperforms the *dependency* model, we only report results for the first. Table 1 presents the coverage of each system, based on the full dataset. Our first result is that type-based cross-lingual bracketing outperforms token-based and achieves up to 91.2% in coverage. As expected, the system of Ziering and Van der Plas (2014) does not cover more than 48.1%. The χ^2 method and the back-off models cover all 3NCs in our dataset. The fact that AWDB$_{type}$ misses 8.8% of the dataset is mainly due to equal distances between aligned words (e.g., *crisis resolution mechanism* is only aligned to closed compounds, such as the Swedish *krislösningsmekanism* or to nouns separated by one preposition, such as the Spanish *mecanismo de resolución de crisis*). In future work, we will add more languages in the hope to find more variation and thus get an even higher coverage.

System	Acc$_{com}$	$harmonic(com)$	com
AWDB$_{token}$ / AWDB$_{type}$	94.4% / 94.4%	91.0% / **92.8%**	270
Zier.v.d.Plas14$_{token}$ / Zier.v.d.Plas14$_{type}$	**87.8%** / 87.2%	44.6% / **62.0%**	180
AWDB$_{type}$	**94.6%**†	**92.9%**†	184
Zier.v.d.Plas14$_{type}$	86.4%	61.8%	
AWDB$_{type}$	**94.1%**†	92.6%	273
χ^2	87.9%	**93.6%**	
AWDB$_{type} \rightarrow \chi^2$	**93.5%**†	**96.6%**†	278
Zier.v.d.Plas14$_{type} \rightarrow \chi^2$	86.7%	92.9%	
χ^2	87.4%	93.3%	
LEFT baseline	80.9%	89.4%	

Table 2: Direct comparison on common test sets; † = significantly better than the systems in comparison

Table 2 directly compares the systems on common subsets (com), i.e., on 3NCs for which all systems in the set provide a result. The main reason why cross-lingual systems make bracketing errors is

the quality of automatic word alignment. AWDB outperforms Ziering and Van der Plas (2014) significantly[7]. This can be explained with the flexible structure of AWDB, which can exploit more data and is thus more robust to word alignment errors. AWDB significantly outperforms χ^2 in accuracy but is inferior in $harmonic(com)$. The last four lines of Table 2 show all systems with full coverage. AWDB's back-off model achieves the best $harmonic(com)$ with **96.6%** and an accuracy comparable to human performance. For AWDB, types and tokens show the same accuracy. The harmonic mean numbers for the system of Ziering and Van der Plas (2014) illustrate that coverage gain of types outweighs a higher accuracy of tokens. Our intuition that token-based approaches are superior in accuracy is hardly reflected in the present results. We believe that this is due to the domain-specificity of Europarl. There are only few instances, where the bracketing of a 3NC type differs from token to token. We expect to see a larger difference for general domain parallel corpora.

Language	**Acc**$_{com}$	**Coverage**	$harmonic(com)$	com
Romance	86.6%	**86.2%**	**86.4%**	201
Germanic	**94.0%**	68.0%	78.9%	

Table 3: Evaluation of language families for AWDB$_{type}$

Table 3 shows the contribution of the Romance (i.e., French, Italian, Portuguese and Spanish) and Germanic (i.e., Danish, Dutch, German and Swedish) language families for AWDB$_{type}$. Romance languages have a higher coverage than Germanic languages. This is because many 3NCs are aligned to a closed Germanic compound, which gives no information on the internal structure. Since Romance languages use open compounds, coverage is higher. On the other hand, Romance languages are worse in accuracy. One reason for this is that they can also produce constructions that violate Behaghel (1909)'s First Law, e.g., *state health service* can be translated to the Portuguese *serviços de saúde estatais* (lit.: *[service of health] state*$_{adj}$). While Ziering and Van der Plas (2014) excluded the pattern NOUN + PREP + NOUN + ADJ, we observed that excluding results with this pattern worsens the overall performance of AWDB. Test set instances with this pattern in any Romance language have significantly[8] more LEFT labels than the total test set. Furthermore, many instances of these cases can be disambiguated using morphosyntactic information such as number, e.g., *world fishing quotas* aligned to the French *quota**s** de pêche mondi**aux*** (*quotas*$_{pl}$ *of fishing*$_{sg}$ *world*$_{adj,pl}$).

As a result, we have 13,622 3NC tokens in context annotated with bracketing structures that are comparable to human annotation. The manual annotation by Vadas and Curran (2007a) resulted in 5582 three-word NPs, that were successfully used to train supervised learners.

5 Conclusion

In this paper, we presented a method for the automatic bracketing of k-partite noun compounds by using the surface structure (i.e., various word positions) in parallel translations as supervision. In an experiment, we extracted 3NCs from a noun compound database comprising ten languages derived from a parallel corpus. Our bracketing system outperforms all systems in comparison with an accuracy of 94.6% and is comparable with human performance.

In future work, we will investigate how to combine partial bracketing results from different languages and how to make the approach independent from parallel data. Along with this paper, we publish[9] the processed dataset and our test set, which can be used as training and test data for supervised learners.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and Alaeddine Haouas for the discussions on French. This research was funded by the German Research Foundation (Collaborative Research Centre 732, Project D11).

[7] Approximate randomization test (Yeh, 2000), $p < 5\%$

[8] z-test for proportions; $p < 5\%$

[9] www.ims.uni-stuttgart.de/data/AWDB.data.tgz

References

Barker, K. (1998). A Trainable Bracketer for Noun Modifiers. In *Canadian Conference on AI*, Volume 1418 of *Lecture Notes in Computer Science*.

Barrière, C. and P. A. Ménard (2014). Multiword Noun Compound Bracketing using Wikipedia. In *ComAComA 2014*.

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*.

Bergsma, S., D. Yarowsky, and K. Church (2011). Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation. In *ACL-HLT 2011*.

Busa, F. and M. Johnston (1996). Cross-Linguistic Semantics for Complex Nominals in the Generative Lexicon. In *AISB Workshop on Multilinguality in the Lexicon*.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*.

Girju, R. (2007). Improving the Interpretation of Noun Phrases with Crosslinguistic Information. In *ACL 2007*.

Kim, S. N. and T. Baldwin (2013). A Lexical Semantic Approach to Interpreting and Bracketing English Noun Compounds. *Natural Language Engineering 19*(3).

Landis, J. R. and G. G. Koch (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics 33*.

Lapata, M. and F. Keller (2004). The Web as a Baseline: Evaluating the Performance of Unsupervised Web-based models for a range of NLP tasks. In *HLT-NAACL 2004*.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph. D. thesis, Macquarie University.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press.

Nakov, P. and M. Hearst (2005). Search Engine Statistics Beyond the N-gram: Application to Noun Compound Bracketing. In *CONLL 2005*.

Pitler, E., S. Bergsma, D. Lin, and K. W. Church (2010). Using Web-scale N-grams to Improve Base NP Parsing Performance. In *COLING 2010*.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *ACL SIGDAT-Workshop*.

Sinha, R. M. K. (2009). Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. In *Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC 2012*.

Tsvetkov, Y. and S. Wintner (2010). Extraction of Multi-word Expressions from Small Parallel Corpora. In *Coling 2010: Posters*, Beijing, China.

Vadas, D. and J. Curran (2007a). Adding Noun Phrase Structure to the Penn Treebank. In *ACL 2007*.

Vadas, D. and J. R. Curran (2007b). Large-scale Supervised Models for Noun Phrase Bracketing. In *PACLING 2007*.

Yeh, A. (2000). More Accurate Tests for the Statistical Significance of Result Differences. In *COLING 2000*.

Ziering, P. and L. Van der Plas (2014). What good are 'Nominalkomposita' for 'noun compounds': Multilingual Extraction and Structure Analysis of Nominal Compositions using Linguistic Restrictors. In *COLING 2014*.

Ziering, P., L. Van der Plas, and H. Schütze (2013). Multilingual Lexicon Bootstrapping - Improving a Lexicon Induction System Using a Parallel Corpus. In *IJCNLP 2013*.