University of Stuttgart
Germany

From a Distance

Using Cross-lingual Word Alignments for Noun Compound Bracketing

Patrick Ziering[♠] Lonneke van der Plas[♢]
[♠]Institute for NLP, University of Stuttgart, Germany
[♢]Institute of Linguistics, University of Malta, Malta
IWCS 2015

Introduction

Structure: Prediction of the bracketing
Semantics: Structure determines the meaning

natural language processing
- LEFT: Processing of **natural languages**
- RIGHT: **Natural processing** of (any) languages

Motivation - Machine Translation:
The correct translation depends on the bracketing

luxury cattle truck $\xrightarrow[\text{to French}]{\text{translation}}$ as:
- LEFT: *camion pour bétail de luxe*
- RIGHT: *camion de luxe pour bétail*

Behaghel's First Law:
Elements which belong close together intellectually will also be placed close together [Beh09]

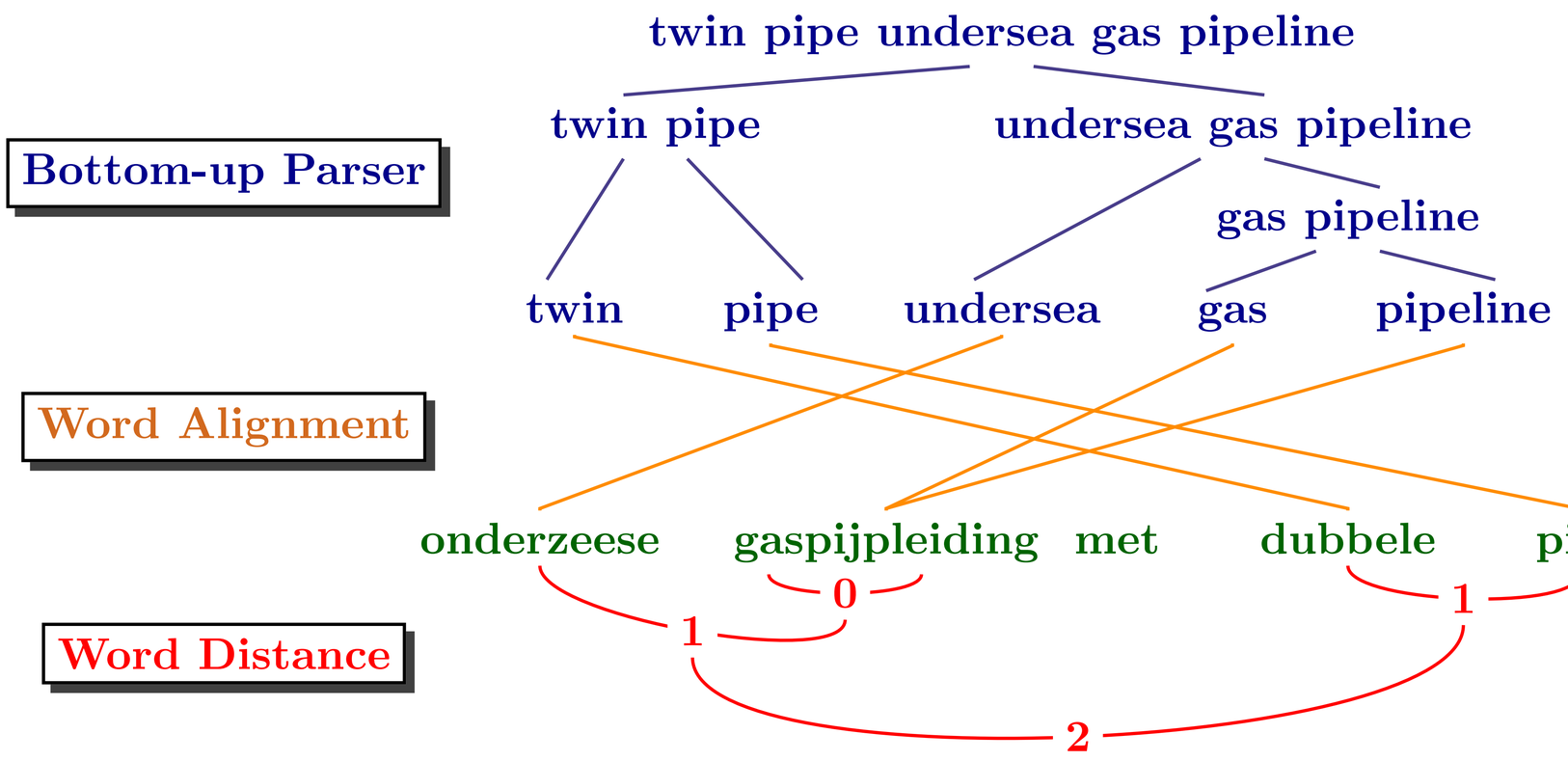
Aligned Word Distance Bracketing (AWDB)

Single Feature: Distance of words aligned to components in a parallel sentence

1: $c_1, \ldots, c_n \Leftarrow N_1, \ldots, N_k$
2: $AW_i \Leftarrow$ set of content words c_i aligns to
3: **while** $|\{c_1, \ldots, c_n\}| > 1$ **do**
4: $(c_m, c_{m+1}) \Leftarrow c$-pair with minimal AWD
5: merge c_m and c_{m+1} to $c_{[m,\ m+1]}$
6: $AW_{[m,\ m+1]} = AW_m \cup AW_{m+1}$
7: **end while**

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

Type-based: Collect all tokens' votes for a type
Final decision: Majority vote across all languages



Benefits of AWDB

- ✓ Unsupervised
- ✓ Accuracy comparable to human performance
- ✓ High coverage
- ✓ Context-based analysis for structural ambiguity
- ✓ Knowledge-lean
- ✓ Language-independent
- ✓ Automatic annotation for supervised learners

Experiments

Tools and data:
- Europarl compound database [ZvdP14]
- 10 languages: *da, de, en, el, es, fr, it, nl, pt, sv*
- Extraction of three-noun compounds (3NCs)
- High-confidence noun filter $P_{noun} = P(noun \mid word)$
  $\Rightarrow$ Final dataset: 14,941 tokens and 8824 types

Human annotation:
- Agreement rate: 90.3%
- $\kappa = 0.717$
- Consensus as test set: 278 tokens (248 types)

Systems in comparison:
- Pattern-based 3NC bracketing [ZvdP14]
- Statistical N-gram approach based on χ^2 [NH05] $\rightarrow$ adjacency model [Mar80]
- Back-off models for [ZvdP14] and AWDB
- LEFT-class baseline

Results (Coverage)

System	Coverage
AWDB *token*/*type*	87.9% / **91.2%**
$\text{AWDB}_{type} \rightarrow \chi^2$	100%
χ^2	100%
[ZvdP14] *token*/*type*	29.9% / **48.1%**
$[\text{ZvdP14}]_{type} \rightarrow \chi^2$	100%
LEFT-class baseline	100%

- Based on full dataset
- Types perform better than tokens
- χ^2 and back-off models cover all 3NCs

Results (Accuracy and Language Families)

System	Acc_{com}	harmonic_{com}	*com*
AWDB *token*/*type*	94.4% / 94.4%	91.0% / **92.8%**	270
[ZvdP14] *token*/*type*	**87.8%** / 87.2%	44.6% / **62.0%**	180
AWDB_{type}	**94.6%**†	**92.9%**†	184
$[\text{ZvdP14}]_{type}$	86.4%	61.8%	
AWDB_{type}	**94.1%**†	92.6%	273
χ^2	87.9%	**93.6%**	
$\text{AWDB}_{type} \rightarrow \chi^2$	**93.5%**†	**96.6%**†	278
$[\text{ZvdP14}]_{type} \rightarrow \chi^2$	86.7%	92.9%	
χ^2	87.4%	93.3%	
LEFT baseline	80.9%	89.4%	

- AWDB outperforms [ZvdP14] and χ^2 significantly in Acc
- $\text{AWDB}_{type} \rightarrow \chi^2$ achieves the best harmonic_{com}

Language	Acc_{com}	Coverage	harmonic_{com}	*com*
Romance	86.6%	**86.2%**	**86.4%**	201
Germanic	**94.0%**	68.0%	78.9%	

References

[Beh09] Otto Behaghel.
Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern.
Indogermanische Forschungen, page 110–142, 1909.

[Mar80] Mitchell Marcus.
A Theory of Syntactic Recognition for Natural Language.
MIT Press, 1980.

[NH05] Preslav Nakov and Marti Hearst.
Search Engine Statistics Beyond the N-gram: Application to Noun Compound Bracketing.
In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CoNLL 2005, pages 17–24, 2005.

[ZvdP14] Patrick Ziering and Lonneke van der Plas.
What good are 'Nominalkomposita' for 'noun compounds': Multilingual Extraction and Structure Analysis of Nominal Compositions using Linguistic Restrictors.
In *Proceedings of COLING 2014*, pages 1047–1058, 2014.

Supported and funded by the Deutsche Forschungsgemeinschaft (DFG SFB 732-D11)

http://www.ims.uni-stuttgart.de