

Word-level Alignment for Multilingual Resource Acquisition

Adam Lopez (alopez@umiacs.umd.edu)*

Michael Nossal (nossal@umiacs.umd.edu)*

Rebecca Hwa (hwa@umiacs.umd.edu)*

Philip Resnik (resnik@umiacs.umd.edu)*†

*University of Maryland Institute for Advanced Computer Studies

†University of Maryland Department of Linguistics

The Treebank Bottleneck

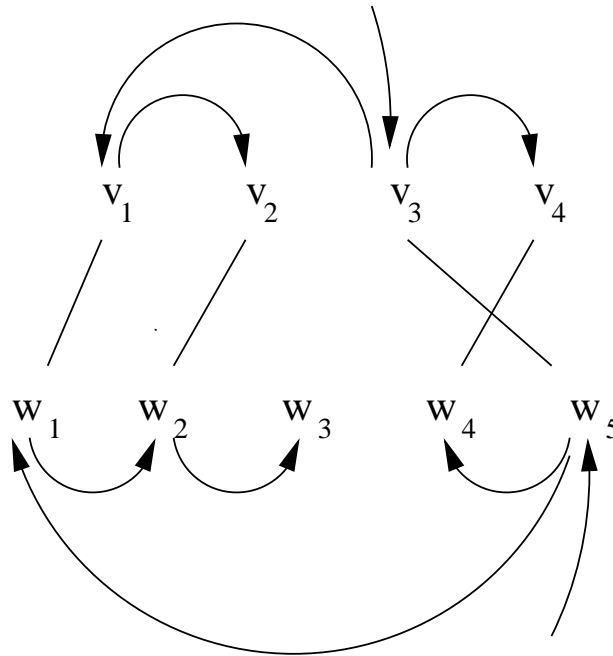
- Stochastic treebank parsers are very accurate (e.g. Charniak, NAACL 2000).
- Creating a treebank is time-consuming and expensive; therefore they are currently available in only a few languages such as English (e.g. Penn Treebank).
- English treebanks could be leveraged to create treebanks in other languages...
 - If the Direct Correspondence Assumption holds (Hwa et. al., ACL 2002).
 - If the available parsers and aligners are good enough.
 - English parsers are very good.
 - We focus on alignment.
 - If robust training algorithms can be found (a la Yarowsky & Ngai, NAACL 2001).

Alignment Algorithm Desiderata

- Improved alignment through use of syntactic knowledge.
- Improved syntactic output through better alignment.
- These goals are complementary.
- Synchronous parsing is a theory which encapsulates these goals.

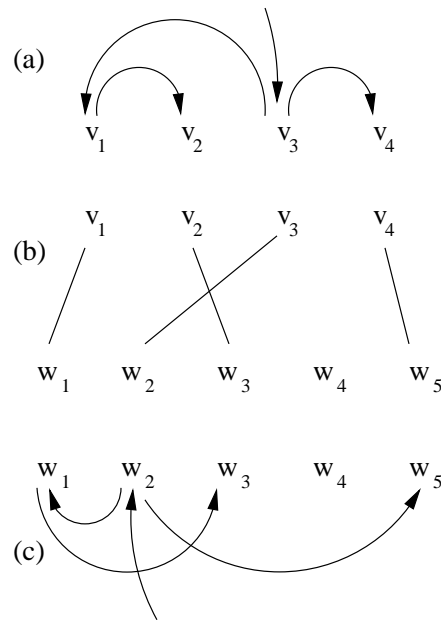
Synchronous Parsing

Models the translation process as dual sentence generation in which a word and its translation in the other sentence are generated in lockstep.



Synchronous Parsing

If we attempt to combine syntax with the output of an aligner that is not sensitive to the constraints imposed by the syntax, the result may be badly formed.



Our Algorithm

- We use a dynamic programming algorithm for synchronous dependency parsing found in Alshawi & Douglas (2000).
- We require output to conform to an input English parse.
- Arbitrary ranking function can be used (we use ϕ^2).
- We only allow 1-to-1 and 1-to-0 word alignments.
- Our handling of null alignments is slightly different.
- Incorporates a search heuristic.
 - Search of entire space would require $O(nm^6)$ steps.
 - Ours search requires $O(nm^3)$ steps.

Experimental Data Set

- Training corpus
 - 56,000 sentence pairs from the Hong Kong News parallel corpus.
 - No annotation required.
- Development set
 - Forty-seven sentences of 25 words or less from sections 001-015 of Chinese Treebank, manually translated to English.
 - English sentences parsed with Collins (ACL, 1999).
 - Context-free parses converted to dependency parses by hand.
 - Manually aligned.
- Test set (46 sentences obtained in similar fashion)

Experimental Setup

- Measured alignment accuracy and Chinese tree accuracy.
- Compared variations on synchronous parsing algorithm.
 - Synchronous parsing without input parses (sim-Alshawi).
 - Synchronous parsing with input English parse.
 - Synchronous parsing with Chinese bigrams.
 - Synchronous parsing with alignment scores initialized from Giza++ alignments.
- Compared with several baselines.

Baseline Results

Baseline Method	AP	AR	AF	CTP
Same Order Alignment	15.7	14.1	14.8	NA
Random Alignment (avg scores)	7.8	7.0	7.4	NA
Forward-chain	NA	NA	NA	37.3
Backward-chain	NA	NA	NA	12.9
Giza++	68.7	40.9	51.3	NA
Hwa, et.al., 2002	NA	NA	NA	44.1

AP = Alignment Precision. AR = Alignment Recall. AF = Alignment F-Score. CTP = Chinese Tree Precision.
All scores are reported as percentages of 100.

Synchronous Parsing Results

Synchronous Parsing Method	AP	AR	AF	CTP
sim-Alshawi (ϕ_A^2)	40.6	36.5	38.4	18.5
sim-Alshawi (ϕ_A^2) + Eng. parse	43.8	39.3	41.4	39.9
sim-Alshawi (ϕ_A^2) + Eng. parse + Ch. bigrams	42.9	38.5	40.6	39.4
sim-Alshawi (ϕ_A^2) + both bigrams	41.5	37.3	39.3	16.5
Giza++ initialization (ϕ_G^2)	51.2	45.9	48.4	11.6
Giza++ initialization (ϕ_G^2) + Engl. parse	49.6	44.6	47.0	44.7

AP = Alignment Precision. AR = Alignment Recall. AF = Alignment F-Score. CTP = Chinese Tree Precision.
All scores are reported as percentages of 100.

Future Directions

- Different scoring functions (e.g. probabilistic).
- Iteration.
- More linguistic knowledge.
- Filtering training data.
- Robust training algorithms for syntax.