NP alignment in bilingual corpora

Gábor Recski, András Rung, Attila Zséder, András Kornai

Computer and Automation Research Institute, Hungarian Academy of Sciences
{recski,zseder,kornai}@sztaki.hu, runga@mokk.bme.hu

Abstract

We created a simple gold standard for English-Hungarian NP-level alignment, Orwell's *1984* by manually verifying the automatically generated NP chunking and manually aligning the maximal NPs and PPs. Since the results are highly impacted by the quality of the NP chunking, we tested our alignment algorithms both with real world (machine obtained) chunkings, where results are in the .35 range for the baseline algorithm which propagates GIZA++ word alignments to the NP level, and on the gold chunkings, where the baseline reaches .4 and our current system reaches .74.

1. Introduction

Aligning the NPs of parallel corpora is logically halfway between the sentence- and word-alignment tasks that occupy much of the MT literature (Gale and Church, 1993; Brown et al., 1993), but has received far less attention (Kupiec, 1993). NP alignment is a challenging problem, capable of rapidly exposing flaws both in the word-alignment and in the NP chunking algorithms one may bring to bear. It is also a very rewarding problem in that NPs are semantically natural translation units, which means that (i) word alignments will cross NP boundaries only exceptionally, and (ii) within sentences already aligned, the proportion of 1-1 alignments will be higher for NPs than words.

Since parallel corpora aligned at the NP level would be an important resource in training and testing performance not just on the NP alignment task itself but also on a range of important tasks already in the focus of MT work, such as factored language modeling (Bilmes and Kirchhoff, 2003), exploration of verbal argument structure (Carreras and Marquez, 2005), and automatically deriving valency dictionaries (Brent and Berwick, 1991), we endeavored to create a simple gold standard for English-Hungarian. Our choice of primary text is Orwell's *1984*, since this already exists in manually verified POS-tagged format in many languages thanks to the Multex (Ide and Véronis, 1994) and Multex East (Erjavec, 2004) projects. The POS-tagged version already catalyzed the development of fully parsed, Penn or Prague Treebank-style, versions for Hungarian, Slovene, Czech, Estonian and quite possibly others we are not aware of (Csendes et al., 2005; Dzeroski et al., 2006; Tadic, 2007), and it is a trivial matter to reformat these as NP-level (CoNLL or Start/End style) annotated text.

Since no English gold standard exists, our first task was to run the text through three independent parsers and NP chunkers (Kudo and Matsumoto, 2001a; Klein and Manning, 2003; Recski and Varga, 2010) and establish a starting point by simple majority vote. Discrepancies between the machine outputs were resolved manually, the fully chunked English and Hungarian texts are available at `http://mokk.bme.hu/multi`the project website at http://mokk.bme.hu/multi. Needless to say, the main interest is not with this largely manual work, but rather with the automated NP alignment process to which we turn now.

2. Alignment

Aligning the English and Hungarian NPs requires some preparation. Koehn (Koehn and Knight, 2003) already merges the NP and PP categories, and we follow this practice because English PPs are cross-linguistically case-marked NPs. Note that our alignment targets are the maximal NPs rather than the minimal (base level) NPs because the highest NP is the one required for factoring the translation process into the translation of predicate/argument structure on the one hand and the translation of NPs on the other.

Table 1 compares three taggers, `yamcha` (Kudo and Matsumoto, 2001b), `mallet` (McCallum, 2002), and `hunchunk` (Recski and Varga, 2009)(Recski and Varga, 2010). All three perform well (over the 94% level) on the standard Penn Treebank NP chunking task (Tjong Kim Sang and Buchholz, 2000) which involves base NPs. Since errors made on the identification of base level NPs percolate up to the analysis of maximal NPs, performance on the maxNP task is not nearly as good (in the 70s) both on the Penn Treebank and on *1984*. We note that `mallet` stays constant when we move from the Penn Treebank to *1984*, `yamcha` improves, and `hunchunk` loses performance (both precision and recall).

task	**fom**	`yamcha`	`mallet`	`hunchunk`
Penn	precision	73.8	73.5	75.0
	recall	71.9	69.8	73.8
	F	72.9	71.6	74.4
1984	precision	74.0	72.4	70.2
	recall	73.1	70.6	70.9
	F	73.8	71.5	70.6

Table 1: Basic figures of merit on maxNP chunking tasks

When it comes to Hungarian, neither `yamcha` nor `mallet` could be optimized well to the task, since they are orders of magnitude slower to train, and run into memory limitations once we start using the kind of more detailed feature sets which are essential to capture the morphology. Therefore, the results are somewhat worse than those produced by `hunchunk`, indicative of the inherent scaling problems of SVMs, MEMMs, and CRFs.

task	**fom**	`yamcha`	`mallet`	`hunchunk`
1984	precision	82.2	84.9	85.1
	recall	82.4	81.9	84.4
	F	82.3	83.4	84.8

Table 2: Basic figures of merit on Hungarian maxNP chunking

Major algorithmic steps in the alignment phase included the following: (1) Stemming of English and Hungarian text using hunmorph (Trón et al., 2005); (2) Building a probabilistic English-Hungarian dictionary from the sentence-aligned bicorpus Hunglish (Varga et al., 2005), using the algorithm ItCo (based on Melamed 1998 and described in detail in Recski et al 2009); (3) Building a probabilistic English-Hungarian dictionary of function words using the same algorithm. We also Extended the main dictionary with a small set of word pairs obtained from *1984* itself. This addition will include word pairs which are typical of the novel, and with proper noun pairs based on low Levenshtein distance capitalized word pairs from *1984*, e.g. *Oceania/Óceánia*

For an English sentence with a set of NPs E and its Hungarian translation with a set of NPs H, an NP alignment is a subset of the set $E \times H$. For each candidate NP-pair $N \in E$, $M \in H$ we evaluate

$$\sum_{\substack{w \in N \\ q \in M}} \frac{p(w,q)}{|M| + |N|}$$

where the probability $p(w,q)$ that two words are translations of each other is obtained from the probabilistic E-H dictionary. When the value of the above formula exceeds a given threshold t, the NP pair (N, M) is deemed part of the alignment.

Both the dictionary building algorithm and the alignment itself take stemmed text as their input. When evaluating the above formula for some pair of NPs (N, M), we disregard function words in both languages to reduce noise. If the NPs contain function words only, we calculate $p(N, M)$ by looking up word pairs in a dictionary of function words.

3. Results and discussion

Since the input to the alignment step is very noisy, this has a major impact on the alignment itself: obviously if the input on the source (target) side is only correct with probability p (q) we can't expect the whole alignment be better than pq. In Table 3, we present not just actual results but also estimates based on the above formula, which give an idea about the potential of the system given the current limitations of the chunkers.

task	**fom**	yamcha	mallet	hunchunk
baseline	precision	47.6	48.5	47.7
	recall	17.7	17.9	17.9
	F	25.8	26.2	26.0
estimate	pq prec	60.8	61.5	59.7
	pq rec	60.2	57.8	59.8
	pq F	60.5	59.6	59.8
current	precision	44.4	44.6	44.2
	recall	37.4	37.5	38.7
	F	40.6	40.7	41.2

Table 3: Baseline alignment algorithm with different chunkings

As the comparison of the F-scores under the three conditions (baseline algorithm, theoretical limit, and our previous algorithm which was taking conditional probabilities from GIZA++) makes clear, the error pattern of our aligner is inherited from the error pattern of the NP chunkers. High quality NP-level alignment would allow us to factor two major sources of cross-language variation: differences between the source and the target in argument structure and differences in the internal composition of the NPs. The former factor is closely correlated to the feasibility of alignment at the NP level, while the latter impacts only our ability to find the NPs. Here we attempt to explore the relative weight of these factors by testing alignment under the idealized condition when the system receives gold (manually tagged) NPs.

condition	**fom**	baseline	current
gold NPs	prec	48.0	77.8
	rec	34.8	70.6
	F	40.3	74.0

Table 4: NP alignment results assuming perfect NP chunking

As can be seen, the alignment task is still very hard, and we are only halfway toward obtaining good results even on this artificial task. Our current algorithm, which simply thresholds alignment pairs based on the conditional probability mass of NPs collected at the word level, is better than the baseline (which simply uses GIZA++(Och and Ney, 2003) at the word level and propagates these to the phrase level), but still very simple, and we plan on exploring several algorithms, such as giving nominal heads greater weight than dependents, by the time of the meeting.

Acknowledgments

Recski worked on the NP chunker. Rung created the gold tagging and alignment. Zséder worked on the aligner. Kornai advised and wrote the paper. We thank Dániel Varga (Budapest Institute of Technology MOKK Media Research).

4. References

Jeff Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *HLT-NAACL*.

Michael R. Brent and Robert C. Berwick. 1991. Automatic acquisition of subcategorization frames from tagged text. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 342–345, Morristown, NJ, USA. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Xavier Carreras and Lluis Marquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *CoNLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged Treebank. In *Lecture Notes in Computer Science: Text, Speech and Dialogue*, pages 123–131.

Saso Dzeroski, Tomaz Erjavec, Nina Ledinek, Petr Pajas, Zdenek Zabokrtsky, and Andreja Zele. 2006. Towards a Slovene dependency treebank. In *Proceedings of LREC-2006*, Genoa, Italy, May 24-26.

Tomaž Erjavec. 2004. Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, pages 1535–1538. ELRA.

William A. Gale and Kenneth Ward Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Nancy Ide and Jean Véronis. 1994. Multext: Multilingual text tools and corpora. In *Proceedings of the 15th conference on Computational linguistics*, pages 588–592, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *In Proc. of the 41st Annual Meeting of the ACL*, pages 311–318.

Taku Kudo and Yuji Matsumoto. 2001a. Chunking with support vector machines. In *NAACL*.

Taku Kudo and Yuji Matsumoto. 2001b. Chunking with support vector machines. In *Proceedings of NAACL 2001*.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 17–22.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. In *http://mallet.cs.umass.edu*.

Dan Melamed. Empirical methods for exploiting parallel texts.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Gábor Recski and Dániel Varga. 2009. A Hungarian NP Chunker. *The Odd Yearbook*.

Gábor Recski and Dániel Varga. 2010. Magyar fönévi csoportok azonosítása. *Általános Nyelvészeti Tanulmányok*.

G. Recski, D. Varga, A. Zséder, and A. Kornai. 2009. Fonevi csoportok azonosítása magyar-angol párhuzamos korpuszban [Identifying noun phrases in a parallel corpus of English and Hungarian]. *VI. Magyar Számitógépes Nyelvészeti Konferencia [6th Hungarian Conference on Computational Linguistics]*.

Marko Tadic. 2007. Building the Croatian dependency treebank: the initial stages. *Suvremena lingvistika*, 33(63):85–92.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL-2000*, pages 127–132. Lisbon, Portugal.

V. Trón, A. Kornai, G. Gyepesi, L. Németh, P. Halácsy, and D. Varga. 2005. Hunmorph: open source word analysis. In *Proceedings of the Workshop on Software*, pages 77–85. Association for Computational Linguistics.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, pages 590–596, Borovets. Bulgaria.

Appendix: sample from the new corpus

0 [It] 0 was [a bright cold day in April] 1 , and [the clocks] 2 were striking [thirteen] 3 .
[Derült , hideg áprilisi nap] 0 volt , [az órák] 1 éppen [tizenhármat] 2 ütöttek .
1-0 2-1 3-2
1 [Winston Smith] 0 , [his chin] 1 nuzzled [into his breast] 2 [in an effort to escape the vile wind] 3 , slipped quickly [through the glass doors of Victory Mansions] 4 , though not quickly enough to prevent [a swirl of gritty dust] 5 [from entering] 6 [along with him] 7 .
[Winston Smith] 0 , [állát] 1 leszegve , gyorsan besurrant [a Győzelem-tömb üvegajtaján] 2 , hogy megszabaduljon [a gonosz széltől] 3 . De nem tudott olyan gyorsan besurranni , hogy ne törjön be [vele] 4 együtt [egy kavicsos porörvény] 5 .
0-0 1-1 3-3b 4-2 7-4 5-5
2 [The hallway] 0 smelt of [boiled cabbage and old rag mats] 1 .
[Az előcsarnokhoz vezető folyosó] 0 [főtt kelkáposzta és öreg rongy lábtörlők szagát] 1 árasztotta .
0-0 1-1s
3 [At one end of it] 0 [a coloured poster , too large for indoor display] 1 , had been tacked [to the wall] 2 .
[Egyik végén] 0 [egy – épületen belüli elhelyezés céljára túlságosan is nagyméretű – plakát] 1 volt [a falra] 2 szegezve .
0-0 1-1 2-2
4 [It] 0 depicted simply [an enormous face , more than a metre wide] 1 : [the face of a man of about forty -five , with a heavy black moustache and ruggedly handsome features] 2 .
Csak [egy hatalmas arc] 0 volt [látható] 1 [rajta] 2 , [méternél is szélesebb arc] 3 : [egy negyvenöt év körüli , sűrű fekete bajuszos , durva vonású férfi arca] 4 .
0-2 1-0 1-3b 2-4
5 [Winston] 0 made [for the stairs] 1 .
[Winston] 0 egyenesen [a lépcső] 1 felé sietett .
0-0 1-1
6 [It] 0 was no use [trying the lift] 1 .
[A felvonóval] 0 nem volt [érdemes] 1 próbálkozni .
1-0b
7 Even [at the best of times] 0 [it] 1 was seldom working , and [at present] 2 [the electric current] 3 was cut off [during daylight hours] 4 .
Még [a jobb időkben] 0 is ritkán működött , jelenleg meg [az áramszolgáltatás] 1 is szünetelt [a nappali órákban] 2 .
0-0 3-1 4-2