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, (since this already exists in manually verified POS-tagged format in many languages thanks to the Multex and MultexEast project) by manually verifying the automatically generated NP chunking (we used the yamcha, mallet and hunchunk taggers) and manually aligning the maximal NPs and PPs. The maximum NP chunking problem is much harder than base NP chunking, with F-measure in the .7 range (as opposed to over .94 for base NPs). 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 idealized (manually obtained) chunkings, where the baseline reaches .4 and our current system reaches .64.

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 (Ku´piec, 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 (Ibe and Veronis, 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, 2001b; 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.html 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).

<table>
<thead>
<tr>
<th>task</th>
<th>fom</th>
<th>yamcha</th>
<th>mallet</th>
<th>hunchunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn</td>
<td>precision</td>
<td>73.8</td>
<td>73.5</td>
<td>75.0</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>71.9</td>
<td>69.8</td>
<td>73.8</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>72.9</td>
<td>71.6</td>
<td>74.4</td>
</tr>
<tr>
<td>1984</td>
<td>precision</td>
<td>74.0</td>
<td>72.4</td>
<td>70.2</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>73.1</td>
<td>70.6</td>
<td>70.9</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>73.8</td>
<td>71.5</td>
<td>70.6</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>task</th>
<th>1984 precision</th>
<th>recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>yamcha</td>
<td>82.2</td>
<td>84.9</td>
<td>85.1</td>
</tr>
<tr>
<td>mallet</td>
<td>82.4</td>
<td>81.9</td>
<td>84.4</td>
</tr>
<tr>
<td>hunchunk</td>
<td>82.3</td>
<td>83.4</td>
<td>84.8</td>
</tr>
</tbody>
</table>

Table 2: Basic figures of merit on Hungarian maxNP chunking

### 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.

<table>
<thead>
<tr>
<th>task</th>
<th>baseline precision</th>
<th>recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>yamcha</td>
<td>47.6</td>
<td>48.5</td>
<td>47.7</td>
</tr>
<tr>
<td>mallet</td>
<td>17.7</td>
<td>17.9</td>
<td>17.9</td>
</tr>
<tr>
<td>hunchunk</td>
<td>25.8</td>
<td>26.2</td>
<td>26.0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>condition</th>
<th>1984 precision</th>
<th>recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>gold NPs</td>
<td>48.0</td>
<td>76.5</td>
<td></td>
</tr>
<tr>
<td>rec</td>
<td>34.8</td>
<td>54.4</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>40.3</td>
<td>63.6</td>
<td></td>
</tr>
</tbody>
</table>

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. Zsédér worked on the aligner. Kornai advised and wrote the paper. We thank Dániel Varga (Budapest Institute of Technology MOKK Media Research).

### 4. References


Philipp Koehn and Kevin Knight. 2003. Feature-rich sta-
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 .

1 [ Derült , hideg áprilisi nap ] 0 volt , [ az órák ] 1 éppen [ tizenhármat ] 2 üöttek .

1-0 2-1 3-2


2-3


0-0 1-3 3b 4-2 7-4 5-5


1 [ Az előcsarnokhoz vezető folyosó ] 0 [ fött kelkáposzta és öreg rongy lábtörők szagát ] 1 árasztotta .

0-0 1-1s

5 [ 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 tüslogosan is nagyméretű – plakát ] 1 volt [ a falra ] 2 szegyezv .

0-0 1-1 2-2

6 [ It ] 0 depicted simply [ an enormous face , more than a metre wide ] 1 [ the face of a man of about forty-live , with a heavy black moustache and ruggedly handsome features ] 2 .

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

8 [ It ] 0 was [ part of the economy drive ] 1 [ in preparation for Hate Week ] 2 .


0-0 1-1 1-2b 2-4b

9 [ The flat ] 0 was [ seven flights ] 1 up , and [ Winston ] 2 , [ who ] 3 was [ thirty -nine ] 4 and had [ a varicose ulcer ] 5 above [ his right ankle ] 6 , went slowly , resting [ several ] 7 times [ ] 7 on the way ] 8 .

10 [ On each landing ] 0 , [ opposite the lift -shaft ] 1 , [ the poster with the enormous face ] 2 gazed [ from the wall ] 3 .


1-0 1-2 2-4 7

11 [ It ] 0 was [ one of those pictures ] 1 [ which ] 2 are so contrived that [ the eyes ] 3 follow [ you ] 4 about when [ you ] 5 move .


1-0

12 [ Big Brother ] 0 is watching [ you ] 1 , [ the caption beneath it ran ] 2 .


0-0 2-2b 2-3

13 [ Inside the flat ] 0 [ a fruity voice ] 1 was reading out [ a list of figures ] 2 [ which ] 3 had [ something ] 4 to do with [ the production ] 5 of [ pig -iron ] 6 .


0-0 1-1 2-2

14 [ The voice ] 0 came [ from an oblong metal plaque like a dulled mirror ] 1 [ which ] 2 formed [ part ] 3 of [ the surface ] 4 of [ the right -hand wall ] 5 .
A hang | 0 | [ a hymnolianos tükörhöz hasonló , téglalap alakú fémlémezóból ] | 1 | áradt , | amely | 2 | [ a jobb kéz felőli falba ] | 3 | volt beépítve .
0-0 1-1
15 | Winston | 0 | turned | [ a switch ] | 1 | and | [ the voice ] | 2 | sank | [ somewhat ] | 3 , | though | [ the words ] | 4 | were | still | distinguishable .
0-0 1-1 2-2 3-3 4-5
16 | The instrument | 0 | ( [ the telescreen ] | 1 , | it | 2 | was | called ) | could | be | dimmed , | but | there | 3 | was | no way of shutting it off completely | 4 .
17 | He | 0 | moved | over | to | the window | 1 : | a smallish , frail figure | 2 , | the meagreness of his body | 3 merely emphasized | by | the blue overalls | 4 | which | 5 | were | the uniform | of | the Party | 7 .
18 | His hair | 0 | was | very fair , | [ his face naturally sanguine ] | 1 , | his skin roughened | 2 | by | coarse soap and blunt razor blades | 3 and | the cold of the winter | 4 | that | 5 | had | just ended .
0-0 1-1 2-2 3-3 4-4
19 Outside , | even | [ through the shut window -pane ] | 0 , | [ the world ] | 1 | looked cold .
[ Okdánnin , még | [ a bezárt ablakon ] | 0 | keresztül is , | [ hidegnek ] | 1 | látszott | [ a világ ] | 2 .
0-0 1-2
20 Down | in | the street | 0 | [ little eddies of wind ] | 1 | were | whirling | [ dust and torn paper ] | 2 | into | spirals | 3 , | and | though | [ the sun ] | 4 | was | shining and | [ the sky ] | 5 | [ a harsh blue ] | 6 , | there | 7 | seemed | to | be | no colour in anything | 8 , | except | the posters | 9 | that | 10 | were | plastered everywhere .
0-8 1-0 2-1 3-2 4-3 5-4 6-5 7-8 9-6
21 | The blackmoustachio ’ s face | 0 | gazed down | from | every commanding corner | 1 .
[ A fekete bajuszos arc ] | 0 | [ minden sarkon ] | 1 | ott | mereszette | [ a szemét ] | 2 .
0-0 1-1
22 | There | 0 | was | one | 1 | [ on the house -front immediately opposite ] | 2 .
Ott | volt | [ a szemközti álló ház homlokzatán ] | 0 | is .

2-0
23 | Big Brother | 0 | is | watching | [ you ] | 1 , | [ the caption ] | 2 said , | while | [ the dark eyes ] | 3 looked deep | [ into Winston ’ s own ] | 4 .
0-0 2-2 3-3 4-4
24 Down | [ at streetlevel ] | 0 | [ another poster , torn at one corner ] | 1 , | flapped fitfully | [ in the wind ] | 2 , | alternately covering and uncovering | [ the single word Ingsoc ] | 3 .
0-0 1-1 2-2 3-3 4-4
25 | The patrols | 0 | did not matter , however .
0-0 1-1 2-2 3-3 4-4 5-5
26 | It | 0 | was | [ the police patrol ] | 1 , | snooping | [ into people ’ s windows ] | 2 .
1-0 2-2 4
27 | The patrols | 0 | did not matter , however .
0-0
28 Only | [ the Thought Police ] | 0 | mattered .
[ Rosszat ] | 0 | csak | [ a Gondolatrendőrség ] | 1 | jelentett .
0-1
29 | Behind Winston ’ s back | 0 | [ the voice from the tele- screen ] | 1 | was | still | babbling away | [ about pig -iron and the overfulfillment of the Ninth Three -Year Plan ] | 2 .
0-0 1-1 2-2
30 | The telescreen | 0 | received and transmitted simultaneously .
0-0