In Search of the Best Method for Sentence Alignment in Parallel Texts

Alexandr Rosen

alexandr.rosen@ff.cuni.cz

Institute of Theoretical and Computational Linguistics
Faculty of Arts and Philosophy, Charles University
Prague, Czech Republic

slovko 2005, 10 November 2005

Plan of the talk:

1. *InterCorp* — a parallel corpus project for 20+ languages
2. Options for sentence alignment
3. Comparison of alignment tools
4. Joining forces
5. Conclusions

1 The project *InterCorp*

`https://trnka.ff.cuni.cz/ucnk/intercorp/`
– sorry, so far only *česky* (in Czech)

Participants:

(mostly from the Charles University's Faculty of Arts and Philosophy)

- Foreign language departments
- Institute of the Czech National Corpus
- Institute of Theoretical and Computational Linguistics

Other projects involving many languages:

- OPUS – Tiedemann & Nygaard (2004) (`http://logos.uio.no/opus/`)
- Joint Research Centre of the European Commision: Acquis Communautaire (`http://www.cba.muni.cz/~zizka/Langtech/`)
- ...

Initial stage:

- Some participants already have some parallel corpora
- They use *ParaConc* for segmentation, alignment and search (`http://www.athel.com/para.html`)

Goals:

- 20+ parallel subcorpora
- Czech always L1 = the pivot
- Balance: mainly fiction, preference for Czech originals
- Alignment: by sentences, as complete and correct as possible
- Distributed pre-processing
- All subcorpora integrated into a single shared resource

For:

- Comparative studies
- Teaching
- Lexicography (including term extraction)
- Extraction of Translation Memories
- Translators
- General public
- ...

2 Alignment

- A parallel corpus is only as good as its alignment.
- Good results of automatic alignment can save manual work.
- Is there a single best all-purpose way to sentence alignment?
- NO! — At least according to previous evaluations of sentence aligners:
  - Langlais et al. (1998)
  - Véronis & Langlais (2000)
  - Singh & Husain (2005)

The choice depends on properties of the input and intended use of the output:

- structural distance between the two texts
  (free or literal translation)
- amount of noise
  (omissions, differences in segmentation)
- typological distance between the two languages
- size of the texts
- acceptable error rate
- acceptable amount of manual checking

With no single text type + diverse languages
→ (probably) no universal solution

Automatic alignment with manual checking

- If near-to-perfect alignment is required and manual checking is possible, how can we integrate manual checking with automatic methods?
- Should we aim for maximum precision **and** recall, or – perhaps – sacrifice recall for higher precision?
- If *safe* links have precision near 100%, maybe only *unsafe* links need to be checked.
- Let's see how to reach maximum precision...

3 Comparison

GC – Gale & Church (1993) – matches sentences by their lengths (counted in characters), the texts should be previously aligned by paragraphs; fast, language-universal `http://nl.ijs.si/telri/Vanilla/`

Mmd – Melamed (1997) – uses cognates (puctuation, numbers, similar words) and (optionally) bilingual lexicon `http://nlp.cs.nyu.edu/GMA/`

Mre – Moore (2002) – combines length-based pre-alignment with a stochastic method to derive a bilingual lexicon, used subsequently to align sentences, proposes 1:1 links only `http://research.microsoft.com/research/downloads/default.aspx`

Hun – Varga et al. (2005) – *HunAlign*, combines length- and lexicon-based methods, can extract lexicon from the text as **Mre** does, customizable `http://mokk.bme.hu/resources/hunalign`

Mmd^{+} – same as **Mmd**, with a 106K-entries bilingual lexicon

Mre* – same as **Mre**, with some final and initial word segments stripped

Mre^{+} – same as **Mre**, with more input data (a 106K-entries bilingual dictionary and an English-Czech pre-aligned corpus of 830K/731K words)

Texts used for testing

AC – 46+46 documents from the English-Czech part of Acquis Communautaire (roughly 1%); all noise was retained (omissions, results of different segmentation rules); segments = paragraphs

1984 – George Orwell's novel, English and Czech (result of the project Multext-East)

FR7 – Seven French fiction/essay books + Czech translations

Results were compared with hand-corrected alignment of full texts:

Text	Cz words	L2 words	Cz segments	L2 segments	All links	1:1 links
AC	62,010	74,986	3,025	2,699	2,685	89%
1984	99,099	121,661	6,756	6,741	6,657	97%
FR7	289,003	337,226	21,936	21,746	21,207	95%

Measures for evaluating alignment

$$recall = \frac{correct\ links}{reference\ links}$$

$$precision = \frac{correct\ links}{test\ links}$$

$$F\text{-}measure = 2 \times \frac{recall\ \times\ precision}{recall\ +\ precision}$$

All links ...

	Reference	Test	Correct	Recall	Precision	F-measure
AC						
GC	2700	2683	2225	82.41	82.93	82.67
Mmd$^+$	2700	2686	2492	**92.30**	92.78	**92.54**
Mre	2700	2313	2218	82.15	95.89	88.49
Mre$^+$	2700	2375	2308	85.48	**97.18**	90.96
1984						
GC	6657	6633	6446	96.83	97.18	97.01
Mmd$^+$	6657	6606	6287	94.44	95.17	94.81
Mre	6657	6167	6110	91.78	99.08	95.29
Mre*	6657	6370	6320	94.94	99.22	97.03
Mre$^+$	6657	6441	6402	96.17	**99.39**	97.76
Hun	6657	6689	6535	**98.17**	97.70	**97.93**
F7						
GC	21207	20868	19427	**91.61**	93.09	92.34
Mre	21207	19512	18801	88.65	**96.36**	**92.35**
Mmd	21207	21057	16161	76.21	76.68	76.44

Links 1:1 only ...

	Reference	Test	Correct	Recall	Precision	F-measure
AC						
GC	2391	2248	2156	90.17	95.91	92.95
Mmd$^+$	2391	2354	2304	96.36	**97.88**	**97.11**
Mre	2391	2313	2218	92.76	95.89	94.30
Mre$^+$	2391	2375	2308	**96.53**	97.18	96.85
1984						
GC	6440	6438	6274	97.42	97.45	97.44
Mmd$^+$	6404	6301	6287	97.62	**99.78**	98.69
Mre	6440	6167	6110	94.88	99.08	96.93
Mre*	6440	6370	6320	98.14	99.22	98.67
Mre$^+$	6440	6441	6402	**99.41**	99.39	**99.40**
Hun	6440	6479	6386	99.16	98.56	98.86
F7						
GC	20116	19220	19427	92.62	**96.94**	94.73
Mre	20116	19512	18801	**93.46**	96.36	**94.89**
Mmd	20116	19714	15539	77.25	78.82	78.03

Observations

Ranking for recall (all links)

Rank	AC	1984	F7
1.	92.30 Mmd$^+$	98.17 Hun	91.61 GC
2.	85.48 Mre$^+$	96.83 GC	88.65 Mre
3.	82.41 GC	96.17 Mre$^+$	76.21 Mmd
4.	82.15 Mre	94.94 Mre*	
5.		94.44 Mmd$^+$	
6.		91.78 Mre	

Ranking for precision (all links)

Rank	AC	1984	F7
1.	97.18 Mre^{+}	99.39 Mre^{+}	96.36 Mre
2.	95.89 Mre	99.22 Mre*	93.09 GC
3.	92.78 Mmd^{+}	99.08 Mre	76.68 Mmd
4.	82.93 GC	97.70 Hun	
5.		97.18 GC	
6.		95.17 Mmd^{+}	

Ranking for F-measure (all links)

Rank	AC	1984	F7
1.	92.54 Mmd^{+}	97.93 Hun	92.35 Mre
2.	90.96 Mre^{+}	97.76 Mre^{+}	92.34 GC
3.	88.49 Mre	97.03 Mre*	76.44 Mmd
4.	82.67 GC	97.01 GC	
5.		95.29 Mre	
6.		94.81 Mmd^{+}	

Similar picture for 1:1 pairs, except for recall (of course ...).

- On noisy texts, Mmd and Mre fare better than GC.
- On well-behaved texts, Mre and Mmd tend to show higher precision.
- GC performed surprisingly well on F7 without paragraph boundaries (the hard region was a book).
- Hun scored best in F-measure.
- Mre and Hun can be expected to gain further points with more input data and lemmatization.
- Also Mmd may profit from creating more cognates by lemmatization.

4 Joining forces

- Can we push precision closer to 100%?
- A single text pair can be processed by more than one aligner and a correct link defined as one on which all (or most) aligners agree.
- Intersection of results $\longrightarrow$ smaller, safer set, a decrease in recall, an increase in precision.

Intersecting results on 1984

	Ref.	Test	Correct	Recall	Prec.	F-msr
GC	6657	6633	6446	**96.83**	97.18	97.01
Mmd^+	6657	6606	6287	94.44	95.17	94.81
Mre^+	6657	6441	6402	96.17	99.39	**97.76**
GC/Mmd^+	6657	6279	6254	93.95	99.60	96.69
GC/Mre^+	6657	6354	6348	95.36	**99.91**	97.58
Mmd^+/Mre^+	6657	6130	6114	91.84	99.74	95.63
$GC/Mmd^+/Mre^+$	6657	6095	6089	91.47	99.90	95.50

Intersecting results on F7

	Ref.	Test	Correct	Recall	Prec.	F-msr
GC	21207	20868	19427	**91.61**	93.09	92.34
Mre	21207	19512	18801	88.65	96.36	**92.35**
Mmd	21207	21057	16161	76.21	76.68	76.44
GC/Mre	21207	17728	17661	83.28	**99.62**	90.72

Tuning Mre on 1984

	Ref.	Test	Correct	Recall	Prec.	F-msr
Mre$^+$ 0.5	6657	6441	6402	**96.17**	99.39	**97.76**
Mre$^+$ 0.8	6657	6415	6487	95.94	99.56	97.72
Mre$^+$ 0.95	6657	6366	6344	95.30	99.65	97.43
Mre$^+$ 0.99	6657	6319	6300	94.64	99.70	97.10
GC/Mre$^+$	6657	6354	6348	95.36	**99.91**	97.58

Tuning Mre on F7

	Ref.	Test	Correct	Recall	Prec.	F-msr
Mre 0.5	21207	19512	18801	**88.65**	96.36	**92.35**
Mre 0.8	21207	19089	18531	87.38	97.08	91.97
Mre 0.95	21207	18571	18105	85.37	97.49	91.03
Mre 0.99	21207	17900	17505	82.54	97.79	89.52
GC/Mre	21207	17728	17661	83.28	**99.62**	90.72

- F-measure is always better for an aligner in solo mode (**Mre**$^+$ and **Mre**)
- A tandem always wins in precision
- **Mre** tuned to higher precision still lags behind a tandem

5 Conclusions and future planes

- Alignment depends on properties of the input, alignment methods differ in their sensitivity to such properties. Thus, word-correspondence methods fare better on noisy texts, where sentence-length-based methods give mixed results.
- Lack of linguistic resources (bilingual lexica) need not be an obstacle for the application of lexically-based methods.
- Higher precision can help the human proofreader focus on unsafe links.
- In order to raise precision, sets of links proposed by different aligners can be intersected. This improves precision by 0.5–3.6 percentage points.

Future:

- lemmatization
- meta-aligner

References

Gale, W. A. & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**(1), 75–102.

Langlais, P., Simard, M., & Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 711–717. Association for Computational Linguistics.

Melamed, I. D. (1997). A portable algorithm for mapping bitext correspondence. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 305–312, Somerset, New Jersey. Association for Computational Linguistics.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 135–144, London, UK. Springer-Verlag.

Singh, A. K. & Husain, S. (2005). Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.

Tiedemann, J. & Nygaard, L. (2004). The OPUS corpus – parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Véronis, J. & Langlais, P. (2000). Evaluation of parallel text alignment systems: the arcade project. In J. Véronis, editor, *Parallel text processing: Alignment and use of translation corpora*, pages 369–388. Kluwer Academic Publishers, Dordrecht.

Typeset on 15th November 2005, at 13:26.