Using BabelNet for high quality text alignment

|  |  |  |
| --- | --- | --- |
|  | Andrzej Zydroń XTM InternationalGerrards CrossSL9 8XF, UKazydron@xtm-intl.com |  |

|  |  |
| --- | --- |
|  |  |

Abstract

This paper describes the problems associated with producing high quality segment alignment in real world scenarios and how BabelNet can be used to attain this goal.

1. Introduction

Text segment alignment became a classical computational linguistics topic following the publication of the seminal paper by Gale & Church in 1993. Although the paper is very interesting, it makes an assumption that the source and target texts correspond at least in terms of paragraphs, if not at the sentence level.

1. The real world

In reality the input material for alignment projects is often quite different from the theoretical work of Gale & Church. Typically you can encounter the following issues:

1. The source and target documents are not the same version:
* Additional text may have been added to the source document
* Text may have been removed or changed
* Text may have been reordered and moved around in the source document
1. In an unconstrained translation environment (i.e. where a CAT tool has not been used) the translator can make arbitrary decisions regarding the translation:
* Sentences may be omitted, the translator regarding them as superfluous, or unwarranted in the target language
* Multiple sentences may be rendered as one sentence
* A single sentence may be translated as multiple sentences
* Complete paragraphs may be ‘reworked’ by the translator to provide a clearer interpretation in the target language.
* The translator may decide to completely rework the structure of the original document, especially if this is an unconstrained translation of a Microsoft Word document.

In a real world environment it is not possible to rely on the classical approach: the documents may be too different to provide the basis of a Gale & Church approach to alignment. In my experience this is the case in 90% of real world alignment projects: someone has the translation and a version of the source, but the two do not correspond.

1. The XTM Aligner approach

The only real viable solution to the problem of alignment lies with the use of large scale lexicons such as BabelNet. Nevertheless it is not possible to rely just on standard lexicons. Translators can often use synonyms that do not correspond directly to the source term. In addition there is the problem of grammatical ‘function’ words that are not covered necessarily adequately by lexicons.

BabelNet provides a very effective way of assessing the viability of matching individual segments of text, and also of identifying text that is out of sequence between source and target elements. The newly released XTM Aligner creates a ‘skeleton’ overlay of the text and assesses the degree of change between the source and target versions of the document. It then uses this overlay to commence the alignment of each individual segment, depending on its viability: the segment may be new or modified, or it may have been moved within the document.

1. Languages and scripts

[The XTM Aligner is designed to work with all scripts and languages, including Cyrillic, Greek, Hebrew, Arabic, Chinese, Japanese, Korean and Devanagari. We have currently licensed the following licenses commercially from the BabelNet team:

* 1. AF - Afrikaans
	2. AR - Arabic
	3. AZ - Azerbaijani
	4. BE - Belorussian
	5. BG - Bulgarian
	6. BN - Bengali
	7. CA - Catalan
	8. CS - Czech
	9. DA - Danish
	10. DE - German
	11. EL - Greek
	12. EN - English
	13. ES - Spanish
	14. ET - Estonian
	15. FA - Farsi
	16. FI - Finnish
	17. FR - French
	18. HE - Hebrew
	19. HI - Hindi
	20. HR - Croatian/Bosnian/Serbian
	21. HU - Hungarian
	22. HY - Armenian
	23. ID - Bahasa (Indonesian)
	24. IS - Icelandic
	25. IT - Italian
	26. JA - Japanese
	27. KA - Georgian
	28. KO - Korean
	29. LT - Lithuanian
	30. LV - Latvian
	31. MK - Macedonian
	32. ML - Maltese
	33. MS - Bahasa (Malaysian)
	34. NL - Dutch
	35. NO - Norwegian
	36. PL - Polish
	37. PT - Portuguese
	38. RO - Romanian
	39. RU - Russian
	40. SK - Slovak
	41. SL - Slovenian
	42. SQ - Albanian
	43. SV - Swedish
	44. SW - Swahili (Kiswahili)
	45. TH - Thai
	46. TL - Tagalog
	47. TR - Turkish
	48. UK - Ukrainian
	49. VI - Vietnamese
	50. ZH - Chinese
1. Excel output

The XTM Aligner creates an output file in Microsoft Excel format, including the use of color to provide hints as to the viability of the match. A probability score is also provided using the standard mathematical scoring of ‘0’ for no probability and ‘1’ for total probability.

The Excel file can then be used by a linguist to check and correct if necessary the alignments. Why use Excel? Excel provides a very good and effective way of correcting misaligned segments. Individual cells can be edited, deleted and the subsequent rows moved up or down to realign the remaining part of the document. It is quick and easy to proof and tidy up the alignment if required.

The aligned Excel document can then be uploaded directly into XTM as is. Any source cells without corresponding target text will be ignored.

The XTM Aligner creates two Excel files:

1. One with all source and target segments
2. A second file with ‘.90+’ in the name containing only high probability matches.

The second ‘90+’ file can be used for ‘fast’ alignment where the translator only needs to review and confirm the matches, rather than trying to work on the alignment of the whole two files.



1. Conclusion

The XTM Aligner is designed to cope with the most demanding alignment projects that exist in the ‘real’ world, including all scripts and languages. It attempts to ‘salvage’ and much translation memory as possible from what may otherwise seem an impossible task. BabelNet is a key element of the alignment process and provides a previously impossible level of accuracy.

References

R. Navigli. **[BabelNet and Friends: A manifesto for multilingual semantic processing](http://wwwusers.di.uniroma1.it/~navigli/pubs/IA_2013_Navigli.pdf%22%20%5Ct%20%22paper)**. [*Intelligenza Artificiale*](http://www.iospress.nl/journal/intelligenza-artificiale/), 7(2), pp. 165-181, IOS Press, 2013

R. Navigli and S. Ponzetto. **[BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network](http://wwwusers.di.uniroma1.it/~navigli/pubs/AIJ_2012_Navigli_Ponzetto.pdf%22%20%5Ct%20%22paper)**. [*Artificial Intelligence*](http://www.journals.elsevier.com/artificial-intelligence/), 193, Elsevier, 2012, pp. 217-250.