

Data Preparation Process

To prepare the data, I first pre-processed the raw OANC text to filter out inconsistent and erroneous sentences. This procedure selects only sentences with a length between 3 and 100 words. In addition, the *Biomed* text forms a large portion of the data (~%30) and harms the domain balance of the corpus. (The other dominating text, *Slate*, is a journal that covers a broader domain rather than a specific one like biomedicine.) Also, *Eggan* and *ICIC* contain only a very small excerpt of the text from two specific domains (~%1 together), which is insignificant in balancing the domain. Therefore, I omitted these portions of the corpus in preparing our unlabeled data.

After the pre-processing step, we parsed the remaining sentences using the constituent-based parser (Charniak & Johnson, 2005). During parsing, some of the sentences which could not be parsed are removed. After parsing, according to the POS tags assigned by the parser, a post-processing step filtered out all the sentences without any kind of verb POS tag. Finally, these parses are used as input to the LTH dependency converter (Johansson & Nugues, 2007) and MaltParser (Nivre et al., 2007) to generate the two kinds of dependency parsers for the out-of-domain unlabeled data (as it was done for labeled and in-domain data). The last column of the table shows the final status of the corpus after all the above procedures.

I have used the data in my master dissertation experiments which is yet being evaluated by the examination. A partial use has been reported in a paper (Zadeh Kaljahi, 2010).

Source	Domain	Original Words Count	Final Selected Sentence Count/Ratio	
911 Report	government, technical	281,093	11,519	%3.8
Berlitz	travel guides	1,012,496	38,818	%12.8
Biomed	technical	3,349,714	-	-
Eggan	fiction	61,746	-	-
ICIC	letters	91,318	-	-
OUP	non-fiction	330,524	13,473	%4.5
PLoS	technical	409,280	13,896	%4.5
Slate	journal	4,238,808	174,456	%57
Verbatim	journal	582,384	19,496	%6.4
Web Data	government	1,048,792	33,632	%11
Total		11,406,155	305,290	

Resulting Data

Source	Domain	Original Words Count	Final Selected Sentence Count/Ratio	
911 Report	government, technical	281,093	11,519	%3.8
Berlitz	travel guides	1,012,496	38,818	%12.8
Biomed	technical	3,349,714	-	-
Eggan	fiction	61,746	-	-
ICIC	letters	91,318	-	-
OUP	non-fiction	330,524	13,473	%4.5
PLoS	technical	409,280	13,896	%4.5
Slate	journal	4,238,808	174,456	%57
Verbatim	journal	582,384	19,496	%6.4
Web Data	government	1,048,792	33,632	%11
Total		11,406,155	305,290	

Annotations

Annotation	Malt Parser Input
Format	CoNLL
Description	Word forms and PoS tags by Charniak & Johanson (2005) parser
File Structure	Single file
Sentence Count	305,290
Size	137MB
Annotation	Charniak Parser Input
Format	<s> sentence </s>
Description	Full OANC sentences
File Structure	1 file for all files inside a original OANC folder (72 files in sum)
Sentence Count	363,107 (after filtering of: unparsable, boundary overlaps, formulas, sentence shorter than 3 & longer than 100 words, unpaired parenthesis, multi-byte characters)
Size	46MB
Annotation	Original Charniak Parses
Format	Penn Treebank parse tree
Description	Penn Treebank (constituency) parses
File Structure	According to input (see <i>Charniak Parser Input</i> above)
Sentence Count	363,107
Size	~120MB
Annotation	Filtered Charniak Parses
Format	Penn Treebank parse tree
Description	Penn Treebank (constituency) parses
File Structure	According to input (see <i>Charniak Parser Input</i> above)
Sentence Count	305,290
Size	~110MB

Annotation	Filtered Charniak Parses
Format	CoNLL
Description	Word forms, PoS tags, and parse trees
File Structure	Single file
Sentence Count	305,290
Size	679MB
Annotation	Malt Parses
Format	CoNLL
Description	Dependency heads (word no.) and relations
File Structure	Single file
Sentence Count	305,290
Size	107MB
Annotation	LTH Dependency Converter Output
Format	CoNLL
Description	Dependency heads (word no.) and relations
File Structure	Single file
Sentence Count	305,290
Size	107MB

References

- Charniak, E. & Johnson, M. 2005, 'Coarse-to-fine n-best parsing and MaxEnt discriminative reranking', *In Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, Michigan, USA, pp. 173-180.
- Johansson, R. & Nugues, P. 2007, 'Extended Constituent-to-dependency Conversion for English', *In Proceedings of NODALIDA 2007*, Tartu, Estonia, pp. 105-112.
- Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S. & Marsi, E. 2007, 'MaltParser: A language-independent system for data driven dependency parsing', *Natural Language Engineering*, vol. 13, no. 2, pp. 95-135.
- Zadeh Kaljahi R. S., 2010, "Adapting Self-training for Semantic Role Labeling", *In Proceedings of the ACL 2010 Student Research Workshop*, Uppsala, Sweden, Pages 91-96.