NLP Resources - Corpora

Jun Gong and Daniel Schulman

January 25, 2006

1. Organization of Corpora
   (a) By Media
      i. Text
         A. Example: Brown corpus (discussed in class).
      ii. Speech (with or without transcriptions)
         A. Example: TIMIT [10]
         B. Designed for developing speech recognition.
         C. 630 speakers, each speaking the same 10 sentences.
      iii. Video
   (b) By Language
      i. multilingual parallel corpora
   (c) By Content [10, 6]
      i. Many written corpora are news stories.
      ii. Good spoken collections of conversational speech (Switchboard)
   (d) By Tagging
      i. POS Tagging
      ii. Categorization
         A. Example: RCV1 and RCV2 [14]
         B. Large collection of Reuters news stories.
         C. Hierarchically categorized.
         D. Used for training and testing text classification systems.
      iii. Treebanks
      iv. Annotation Graphs [1]
         A. Represent all corpus annotations as a directed acyclic graph
         B. Intented for text, audio, pos, treebanks, etc.
2. Major Resources
   (a) Linguistic Data Consortium [10]
      i. Academic/Business consortium, led by UPenn

ii. Big collection of corpora, mostly non-free.

(b) Evaluation and Language Resources Distribution Agency (ELDA) [6]

i. Part of the European Languages Resource Association (ELRA)

ii. Mainly multilingual (European languages) corpora

(c) International Computer Archive of Modern and Medieval English (ICAME) [7]

i. English-only (US, UK, historical, others)

ii. Older than other collections (nothing new since 1999?)

iii. Includes Brown corpus

(d) NIST Collection of Reuters Corpora [14, 12]

i. Two large collections (one English, one multilingual) of news stories

ii. Manually categorized

iii. Free for research use

(e) British National Corpus [2]

i. Very large (100 million words) and varied (spoken & written)

ii. Tagging

A. C5 tagset (basic) - entire corpus (automatic tagged)

B. C7 tagset (extended) - 2 million words (manually tagged)

C. Tagged with CLAWS4 tagger [11]

(f) European Corpus Initiative Multilingual Corpus I (ECI/MCI) [5]

i. Large, varied topics and languages (mainly European)

ii. Not free, but cheap (50 euros)

3. Web as Corpus [9]

(a) Really big (estimate 2000 billion words in 2003)

(b) Untagged, but good for word usage statistics

(c) Pages within a site approximate a domain-specific corpus

(d) Multi-language web pages make up a parallel corpus

(e) Issues:

i. Is it representative?

ii. Rates of incorrect words higher than many traditional corpora

iii. Search engines don't return what you want

4. Example Uses

(a) Corpus-Based Stemming [15]

i. Objective: Common stemmers are too agressive. A corpus-based approach improves precision.

ii. Methodology: Modify agressive stemming using a corpus-derived similarity value.

iii. Corpora: WEST legal documents, WSJ(87-91) and WSI(91) from TREC.

(b) Corpus-Based Machine Translation [4]

i. Objective: Machine translation using statistics of a bi-lingual text corpus.

ii. Methodology: Estimate most probable translation of a word with tri-grams.

iii. Corpus: Proceedings of Canadian parliament (100 million words French-English).

iv. Results: 48% acceptable, 5% exactly correct.

(c) Corpus-Based Parsing [13]

i. Objective: A self-learning parser that may extend itself without relying on extra input.

ii. Methodology: Generate hypothesis from partial results - choose the ones generated most.

iii. Corpus: WSJ corpus (for verifying validity).

(d) Corpus-Based Word Sense Disambiguation [8]

i. Objective: A system that learns to disambiguate using an untagged corpus as examples.

ii. Methodology

A. Compute closely-related sentence context from a Machine Readable Dictionary

B. Compare similarities of an appearance of a word with the trained context

iii. Corpus: Treebank-2

iv. Lexicon: WordNet

v. Results: 92% average success rate.

(e) Corpus-Based Tagging [3]

i. Brill tagger - as discussed in class.

References

[1] Bird, S., and Liberman, M. A formal framework for linguistic annotation. *Speech Commun. 33*, 1-2 (2001), 23–60.

[2] http://www.natcorp.ox.ac.uk, Jan 2006.

[3] Brill, E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguist. 21*, 4 (1995), 543–565.

[4] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. A statistical approach to machine translation. *Comput. Linguist. 16*, 2 (1990), 79–85.

[5] http://www.elsnet.org/resources/ecicorpus.html, Jan 2006.

[6] http://www.elda.org, Jan 2006.

[7] http://nora.hd.uib.no/icame.html, Jan 2006.

[8] Karov, Y., and Edelman, S. Similarity-based word sense disambiguation. *Comput. Linguist. 24*, 1 (1998), 41–59.

[9] Kilgarriff, A., and Grefenstette, G. Introduction to the special issue on the web as corpus. *Comput. Linguist. 29*, 3 (2003), 333–347.

[10] http://www.ldc.upenn.edu, Jan 2006.

[11] Leech, G., Garside, R., and Bryant, M. Claws4: the tagging of the british national corpus. In *Proceedings of the 15th conference on Computational linguistics* (Morristown, NJ, USA, 1994), Association for Computational Linguistics, pp. 622–628.

[12] Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res. 5* (2004), 361–397.

[13] Liu, R.-L., and Soo, V.-W. A corpus-based learning technique for building a self-extensible parser. In *Proceedings of the 15th conference on Computational linguistics* (Morristown, NJ, USA, 1994), Association for Computational Linguistics, pp. 441–446.

[14] http://trec.nist.gov/data/reuters/reuters.html, Jan 2006.

[15] Xu, J., and Croft, W. B. Corpus-based stemming using cooccurrence of word variants. *ACM Trans. Inf. Syst. 16*, 1 (1998), 61–81.