Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation

Shane Bergsma **David Yarowsky** **Kenneth Church**
Department of Computer Science and HLTCOE, Johns Hopkins University

Resolving Coordination Ambiguity

Problem: Which words are being linked by a conjunction?
Specifically: Coordination in complex Noun Phrases (NPs)

[dairy and meat] production
asbestos and [polyvinyl chloride]

Interpretation: yes: [dairy production] and [meat production]
no: [asbestos chloride] and [polyvinyl chloride]

Hard problem! Treebank-trained parsers fail on it

Why? Could help Information Retrieval

Could also help syntactic Machine Translation

Like PP-attachment, etc., ***the specific lexical items matter***
- Both sequences above have part-of-speech tags: "*NN and NN NN*"
- We need to **look beyond labeled data** for the key information
  e.g. paraphrase (on the web) diagnostic (*Nakov & Hearst , 2005*)

Size Matters for Resolving Coordination

WSJ portion of Penn Treebank
• 1 **MILLION** words
(labeled data)

Marcus et al. (1993)
We use: *Vadas & Curran (2007)* annotations, giving syntax of all NPs in WSJ portion of Penn Treebank

Bitexts
• 1 **BILLION** words
(bilingual data)

Koehn (2005): Europarl
Callison-Burch et al. (2010): WMT 2010
We use: *English*-to-{*Czech, Danish, German, Greek, Spanish, Finnish, French, Italian, Dutch, Portuguese, Swedish*} bilingual data

Web text (N-grams)
• 1 **TRILLION** words
(unlabeled data)

Brants & Franz (2006): Google N-gram Corpus
We use Google N-grams V2 *(Lin et al., 2010)*

Better Together: Mono and Bilingual Data

Phrase		Corpus Evidence (English and Foreign)	Pattern
dairy (1) and meat (2) production (3)	English	production of dairy and meat	3 of 1 and 2
	English	dairy production and meat production	1 3 and 2 3
	English	meat and dairy production	2 and 1 3
	Spanish	producción láctea y cárnica *(English: production dairy and meat)*	3 1 … 2
	Finnish	maidon- ja lihantuotantoon *(English: dairy- and meatproduction)*	1 - … 2 3
	French	production de produits laitiers et de viande *(English: production of products dairy and of meat)*	3 … 1 … 2
asbestos (1) and polyvinyl (2) chloride (3)	English	polyvinyl chloride and asbestos	2 3 and 1
	English	asbestos , and polyvinyl chloride	1 , … 2 3
	English	asbestos and chloride	1 and 3
	Portuguese	o amianto e o cloreto de polivinilo *(English: the asbestos and the chloride of polyvinyl)*	1 … 3 … 2
	Italian	l' asbesto e il polivinilcloruro *(English: the asbestos and the polyvinylchloride)*	1 … 2 3

Exploiting Unlabeled Text

Given some labeled examples of the two types of coordination:

coal and steel money
insurrection and regime change
North and South Carolina
business and computer science
the Bosporus and Dardanelles straits
pollution and transport safety
rocket and mortar attacks
the environment and air transport

• Create features for **counts** of **patterns** in:
  a) **Bitexts** (Foreign Patterns)
  b) **Web text** (English Patterns)
• Also create **binary features** for words/tags

Bilingual "View"

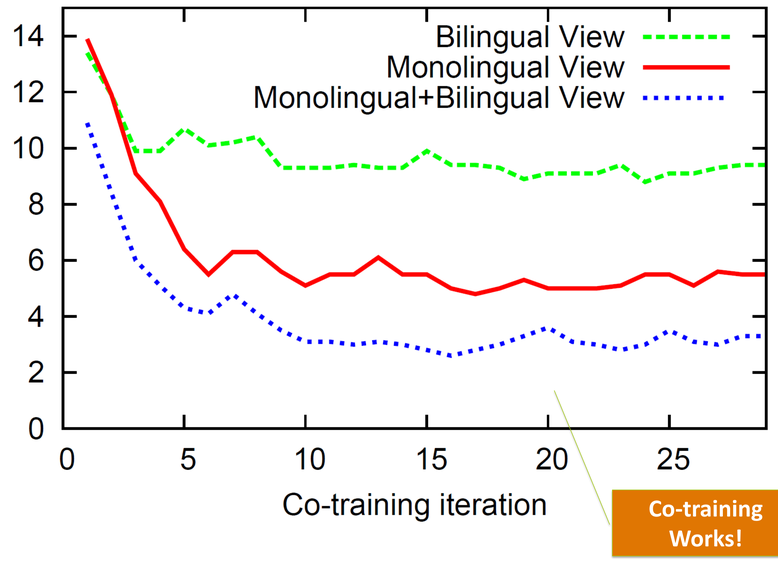
Monolingual "View"

• Train a logistic regression classifier to classify the coordination type

• Iteratively ***co-train*** using the Monoclassifier to label new examples for the bilingual classifier, and vice versa

Results on Europarl Examples

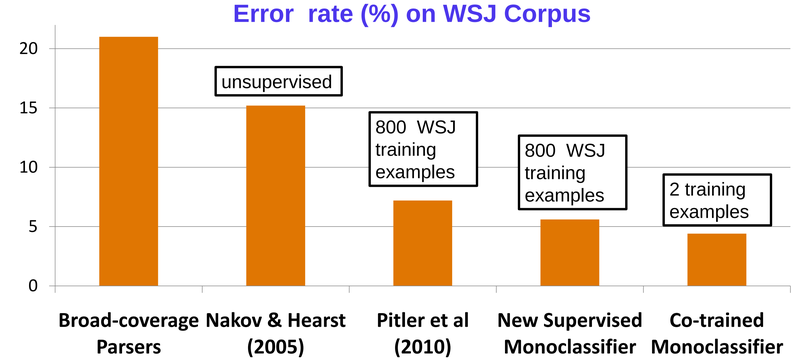
Error rate (%) of Co-trained Classifiers on Europarl test data



Co-training Works!

Results on WSJ Examples

Error rate (%) on WSJ Corpus



• ***Co-trained classifier using only monolingual features advances the state-of-the-art on WSJ data***

Check out our new annotated data and evaluation scripts at:
http://www.clsp.jhu.edu/sbergsma/coordNP.ACL11.zip