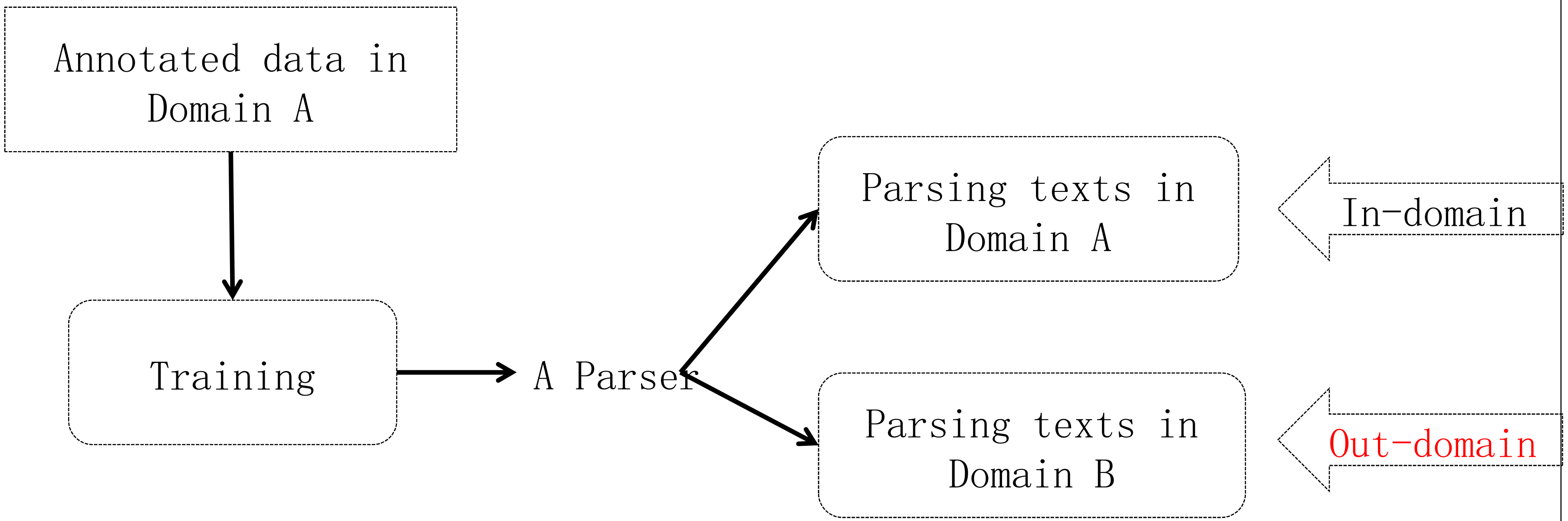
Part C: Parsing the web and domain adaptation

In-domain vs out-domain



Motivation

- Few or no labeled resources exist for parsing text of the target domain.

- Unsupervised grammar induction?
  - Lots of work
  - Accuracies significantly lag behind those of supervised systems
  - Only on short sentences or assuming the existence of gold POS tags

- Build strong parsers by exploring labeled resources of existing domains plus unlabeled data for the target domain.

Outline

- Three shared tasks for parsing out-domain text
- Approaches for parsing out-domain text
  - news domain
  - web data

Shared tasks

- CoNLL 2007 shared task on domain adaptation
- CoNLL 2009 shared task on domain adaptation
- SANCL 2012 parsing the web

CoNLL 2007 shared task on domain adaptation

- Setup for the domain adaptation track
  - Data
    - Train: Large-scale labeled data for the source domain (WSJ)
    - Development: labeled data for biomedical abstracts
    - Test: labeled data for chemical abstracts
    - Unlabeled: large-scale unlabeled data for each train/dev/test.
  - The goal is to use the labeled data of the source domain, plus any unlabeled data, to produce accurate parsers for the target domains.

CoNLL 2009 shared task on domain adaptation

- Setup for the domain adaptation track
  - Czech, German, English (Brown corpus)
  - No unlabeled data
- Provide initial out-of-domain results for the three languages.

SANCL 2012: Parsing the web

- Data Setup (Petrov and McDonald, 2012)
  - Labeled data
    - Train: WSJ-train
    - Development: emails, weblogs, WSJ-dev
    - Test: answers, newsgroups, reviews, WSJ-test
  - Unlabeled data
    - Large-scale unlabeled data for all domains
- The goal is to build a single system that can robustly parse all domains.

Data sets for SANCL 2012

	Training	Development			Evaluation			
	WSJ-train	Emails	Weblogs	WSJ-dev	Answers	Newsgroups	Reviews	WSJ-eval
Sentences	30,060	2,450	1,016	1,336	1,744	1,195	1,906	1,640
Tokens	731,678	29,131	24,025	32,092	28,823	20,651	28,086	35,590
Types	35,933	5,478	4,747	5,889	4,370	4,924	4,797	6,685
OOV	0.0%	30.7%	19.6%	11.8%	27.7%	23.1%	29.5%	11.5%

	Emails	Weblogs	Answers	Newsgroups	Reviews
Sentences	1,194,173	524,834	27,274	1,000,000	1,965,350
Tokens	17,047,731	10,356,284	424,299	18,424,657	29,289,169
Types	221,576	166,515	33,325	357,090	287,575

Approaches for parsing canonical out-domain text (CoNLL07)

- Feature-based approaches
  - Only include features that transfer well (Dredze+, 07)
  - Structural corresponding learning: transform features from source domain to target domain (Shimizu and Nakagawa, 07)
- Ensemble-based approaches
  - Stacking (Dredze+, 07)
  - Co-training (Sagae and Tsujii, 07)
  - Variant of self-training (Watson and Briscoe, 07)

Approaches for parsing canonical out-domain text (CoNLL07)

- Other approaches
  - Tree revision rules for target domain (Attardi+, 07)
  - Training instance weighting (Dredze+, 07)
  - Hybrid: use the output of a Constraint Grammar parser (Bick, 07)
  - Use collocations and relational nouns from unlabeled target domain data (Schneider+, 07)

Frustratingly hard domain adaptation (Dredze+, 2007)

- Theoretical work on domain adaptation attributes adaptation loss to two sources (Ben-David+, 2006)
  - Difference in the distribution between domains
  - Difference in labeling functions
- The error analysis of Dredze+ (2007) suggests that the primary source of errors is the difference in annotation guidelines between treebanks.

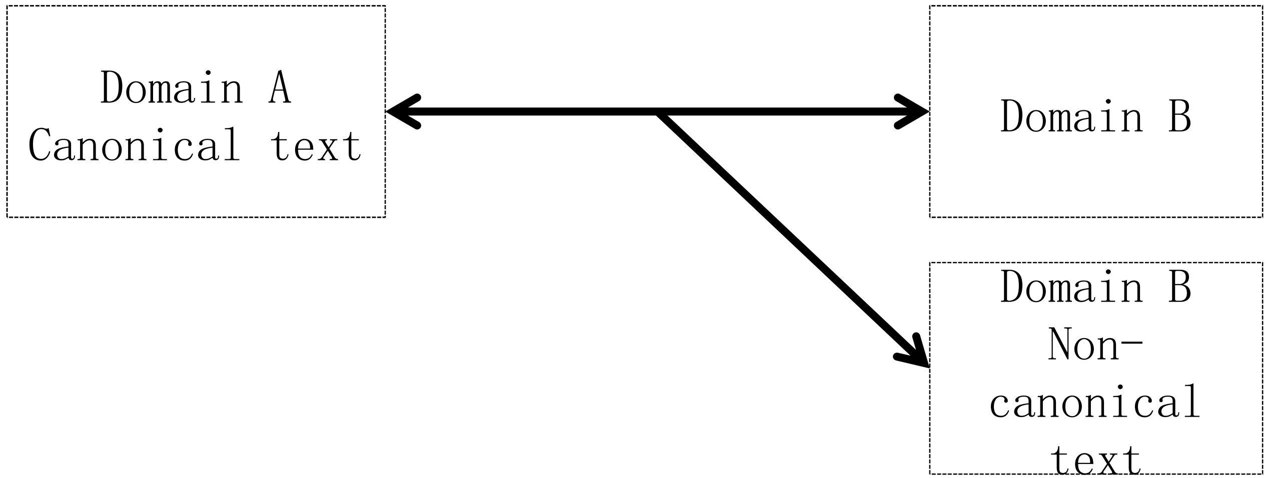
Frustratingly hard domain adaptation (Dredze+, 2007)

- Challenges for adaptation from WSJ (90%) to BIO (84%)
  - Annotation divergences between BIO and WSJ
  - Unlike WSJ, BIO contains many long sequence of digits.
  - Complex noun phrases
  - Appositives
  - WSJ uses fine-grained POS tags such as NNP, while BIO uses NN.
- Long list of failed attempts

Frustratingly hard domain adaptation (Dredze+, 2007)

- Feature manipulation
  - Remove features less likely to transfer
  - Add features more likely to transfer
  - Using word clustering features
- Parser diversity
  - Ensemble of parsers (similar to stacking and bagging)
- Target focused learning
  - Assign higher weights to instances similar to the target when training

Domain + non-canonical text differences



Parsing non-canonical out-domain text (SANCL)

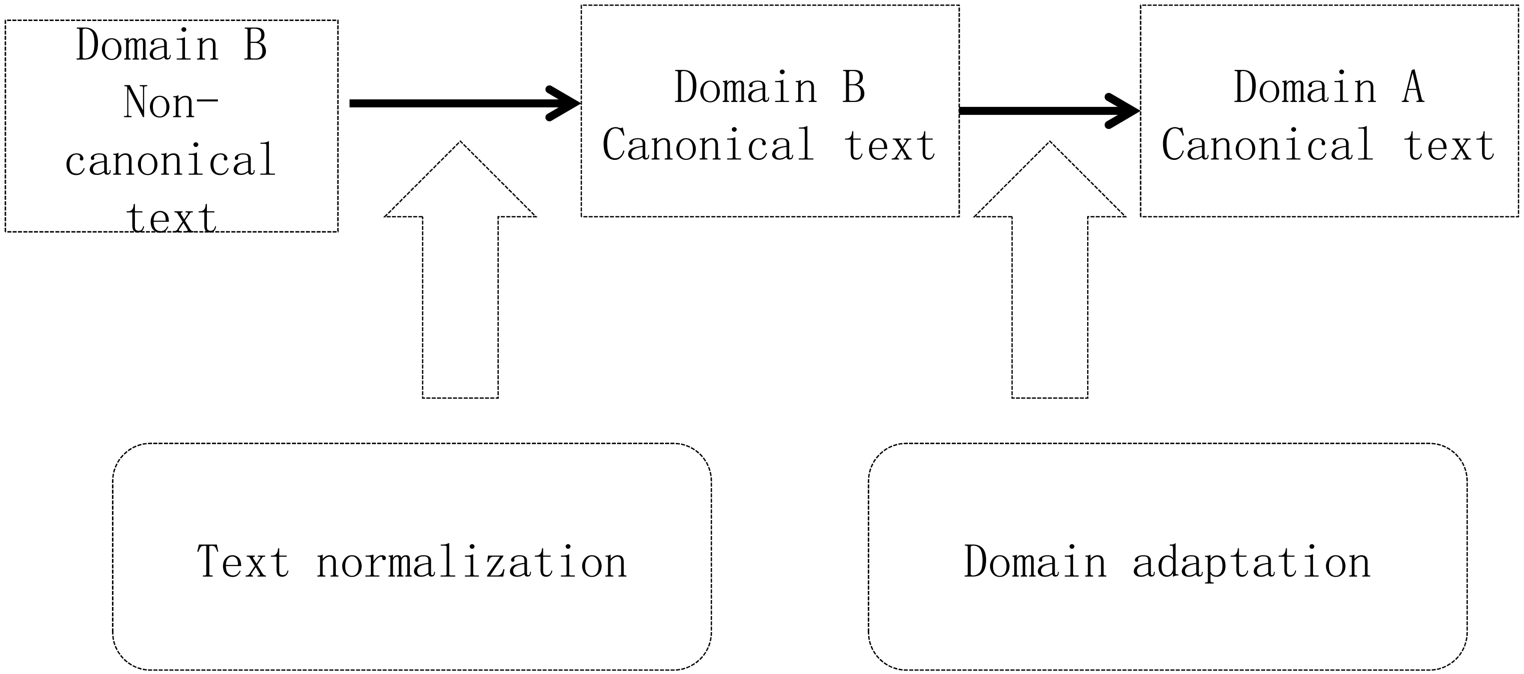
- What is new?
  - Inconsistent usage of punctuation and capitalization
  - Lexical shift due to increased use of slang, technical jargon, or other phenomena.
  - Spelling mistakes and ungrammatical sentences
  - Some syntactic structures are more frequently used in web texts than in newswire
    - Questions, imperatives, long lists of names, sentence fragments…

Examples

- Plz go there.
- I like it very much!!!!!!
- Gooooooooo
- …

Approaches for parsing non-canonical out-domain text (SANCL)

Approaches


Domain B Non-canonical text
Domain B Canonical text
Domain A Canonical text
Text normalization
Domain adaptation

Approaches for parsing non-canonical out-domain text (SANCL)

- Main approaches
  - Text normalization (preprocessing)

  - Ensemble of parsers
  - Self-training for constituent parsing
  - Word clustering/embedding
  - Co/tri-training (unsuccessful)
  - Instance weighting and genre classification

Text normalization

- Preprocessing the data leads to better POS tagging and parsing performance. (Foster, 2010; Gadde+, 2011; Roux and Foster+, 2012)

Text normalization

- The preprocessing rules of (Roux, Foster+, 2012)
  - Emoticon => comma or full stop
  - Email address, URL => generic strings
  - Uppercased words => lowercased
  - Abbreviations, spelling variants (plz, ppl) => standard form
  - nt; s => n't; 's
  - Repeated punctuation (!!!) => collapsed into one
  - List items (# 2) => removed

Text normalization

- The preprocessing rules of (Seddah+, 2012)
  - An Ontonote/PTB token normalization stage
  - Smileys, URLs, email addresses, similar entities
  - Correct tokens or token sequences
    - Spelling error patterns
    - Lowercasing
    - Rewriting rules for dealing with frequent amalgams (*gonna* or *im*)

	Ontonotes dev		e-mail dev		weblog dev	
	all	unk	all	unk	all	unk
−corr	96.5	92.3	88.9	62.4	94.7	87.2
+corr	96.5	92.9	90.4	72.1	94.7	87.3

Text normalization

- The preprocessing rules of (McClosky+, 2012)
  - High-precision text replacements
    - 1,057 spelling auto-correction rules (yuo => you) from Pidgin instant messaging client
    - 151 common Internet abbreviations (LOL => “laughing out loud”)
  - Limited gain
    - Such spelling errors are infrequent in the unlabeled data.

Ensemble of parsers

- Product-of-experts (Alpage, DCU-Paris13)
- Stacking (IMS, Stanford, UPenn)
- Voting (CPH-Trento, DCU-Paris, HIT)
- Bagging (HIT)
- Up-training (IMS)
- Re-ranking (DUC-Paris13, IMS, Stanford)
- Model merging (OHSU, Stanford)

- Obtain large improvement gain.
  - More like improvement in in-domain parsing
  - Contribution to domain adaptation?

Exploring unlabeled data

- Self-training (successful for constituent parsers)
  - Two-stage generative model and reranker (Charniak and Johnson, 2005)
  - Generative PCFG-LA model (Petrov and Klein, 2007)
- Word clusters or embeddings
- Co/tri-training (unsuccessful for dependency parsers)

Why self-training is unsuccessful for dependency parsing?

- Generative models suffer less from the over-fitting problem during training.
- Current dependency parsing models are commonly discriminative.
  - Linear models with online training, no probabilistic explanation.
  - Generative models leads to unsatisfactory accuracy.

Evaluation results

- Top 4 systems of SANCL on POS tagging
  - Tagging performance is very important

Team	Answers	Newsgroups	Reviews	WSJ	Averaged
DCU-Paris (Roux, Foster+)	**91.79**	**93.81**	**93.11**	97.29	**92.90 (1)**
HIT (Zhang+)	90.99	93.32	90.65	**97.76**	91.32 (2)
IMS (Bohnet+)	91.07	91.70	90.01	97.57	90.93 (3)
Stanford (McClosky+)	90.30	91.49	90.46	95.00	90.75 (4)

Which one is the best/most important?

- Main approaches
  - Text normalization (preprocessing)
  - Ensemble of parsers
  - Self-training for constituent parsing
  - Word clustering/embedding
  - Co/tri-training (unsuccessful)
  - Instance weighting and genre classification

End of Part C

References

- G. Attardi, F. Dell'Orletta, M. Simi, A. Chanev, and M. Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using desr. In Proc. of the CoNLL 2007 Shared Task.
- Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In NIPS 2006.
- E. Bick. 2007. Hybrid ways to improve domain independence in an ML dependency parser. In Proc. of the CoNLL 2007 Shared Task.
- Eugene Charniak and Mark Johnson. 2005. Coarse-to-Fine n-Best Parsing andMaxEnt Discriminative Reranking. Proceedings of ACL 2005, pp173–180.
- D. Das and S. Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In Proc. of ACL-HLT.
- M. Dredze, J. Blitzer, P. P. Talukdar, K. Ganchev, J. Graca, and F. Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In Proc. of the CoNLL 2007 Shared Task.
- Matthew S. Dryer and Martin Haspelmath, editors. 2011. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. http://wals.info/.

References

- Jennifer Foster. 2010. “cba to check the spelling” investigating parser performance on discussion forum posts. In Proceedings of HLT NAACL.
- Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. # hardtoparse: Pos tagging and parsing the twitterverse. In Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.
- Phani Gadde, L. V. Subramaniam, and Tanveer A. Faruquie. 2011. Adapting a wsj trained part-of-speech tagger to noisy text: Preliminary results. InProceedings of Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data.
- Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štepánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Proceedings of CONLL 2009 shared task, pp1–18
- Terry Koo, Xavier Carreras, Michael Collins. 2008. Simple semi-supervised dependency parsing. In Proceedings of ACL 2008, pp 595–603

References

- David McClosky, Wanxiang Che, Marta Recasens, Mengqiu Wang, Richard Socher, and Christopher D. Manning. 2012. Stanford's System for Parsing the EnglishWeb. In Notes of the First Workshop on SANCL.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In Proceedings of EMNLP.
- Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In Proceedings of EMNLP-CONLL 2007, pp915–932
- S. Petrov, D. Das, and R. McDonald. 2011. A universa part-of-speech tagset. In ArXiv:1104.2086.
- Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages404–411
- Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 Shared Task on Parsing the Web. In Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL).

References

- Joseph Le Roux, Jennifer Foster, Joachim Wagner, Rasul Samad Zadeh Kaljahi, and Anton Bryl. 2012. Dcuparis13 systems for the sancl 2012 shared task. In Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL).
- K. Sagae and J. Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In Proc. of the CoNLL 2007 Shared Task.
- G. Schneider, K. Kaljurand, F. Rinaldi, and T. Kuhn. 2007. Pro3Gres parser in the CoNLL domain adaptation shared task. In Proc. of the CoNLL 2007 Shared Task.
- Djame Seddah, Benoit Sagot, and Marie Candito. 2012. Robust pre-processing and semi-supervised lexical bridging for user-generated content parsing. In Notes of the First Workshop on SANCL.
- N. Shimizu and H. Nakagawa. 2007. Structural correspondence learning for dependency parsing. In Proc. of the CoNLL 2007 Shared Task.
- Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target Language Adaptation of Discriminative Transfer Parsers. In Proc. of NAACL.
- Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In Proceedings of NAACL-HLT.

References

- R. Watson and T. Briscoe. 2007. Adapting the RASP system for the CoNLL07 domain-adaptation task. In Proc. of the CoNLL 2007 Shared Task.
- Meishan Zhang, Wanxiang Che, Yijia Liu, Zhenghua Li, Ting Liu. 2012. HIT dependency parsing: Bootstrap aggregating heterogeneous parsers. In Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)