It Depends on the Translation:

Unsupervised Dependency Parsing via Word Alignment

Samuel Brody

Department of Biomedical Informatics
Columbia University

samuel.brody@dbmi.columbia.edu

EMNLP 2010

Outline

Parsing - State of Affairs

Supervised

- Posit a grammar
- Train on annotated data (Treebank)
- Apply to target text

Drawbacks

- Which grammar?
- Annotation is expensive
- Strong domain dependence

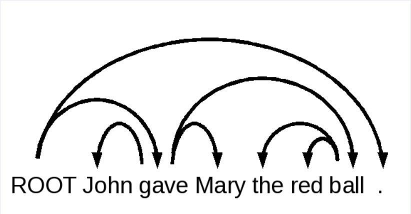
Unsupervised Dependency Parsing

Dependency Parsing

- Simplify by removing latent structure

Unsupervised Learning

- Learn without annotation



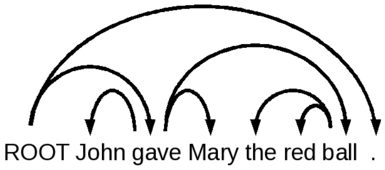
Unsupervised Dependency Parsing

Dependency Parsing

- Simplify by removing latent structure

Unsupervised Learning

- Learn without annotation



Early Work

- Carroll and Charniak (1992) - PCFG over parts-of-speech
- Yuret (1998) - Mutual Information between head & dependent
- Paskin (2001) - Learns $P(dependent|head, direction)$

Dependency Model with Valence

DMV - Klein and Manning (2004)

- Significantly outperformed previous approaches
- First to beat the adjacent-word baseline
- Basis for many recent methods (Cohen and Smith, 2009; Headden III et al., 2009)

Dependency Model with Valence

DMV - Klein and Manning (2004)

- Significantly outperformed previous approaches
- First to beat the adjacent-word baseline
- Basis for many recent methods (Cohen and Smith, 2009; Headden III et al., 2009)

Reasons for Success

- Use of PoS rather than lexical items
- Notion of valence
- Treatment of distance

?

Goal: An Experimental Framework

Goal: An Experimental Framework

Requirements:

Goal: An Experimental Framework

Requirements:

- Modular:
  easily add/remove models

Goal: An Experimental Framework

Requirements:

- Modular:
  easily add/remove models
- Make use of different information
  - PoS
  - Lexical
  - Word Categories

Goal: An Experimental Framework

Requirements:

- Modular:
  easily add/remove models
- Make use of different information
  - PoS
  - Lexical
  - Word Categories

What is a good framework?

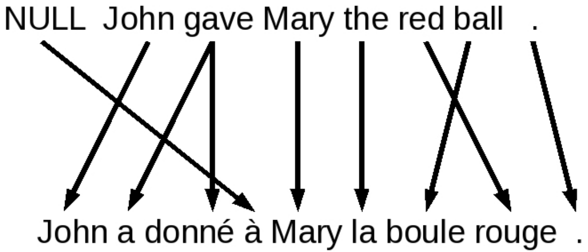
Statistical Machine Translation

The SMT Problem

John gave Mary the red ball .

John a donné à Mary la boule rouge .

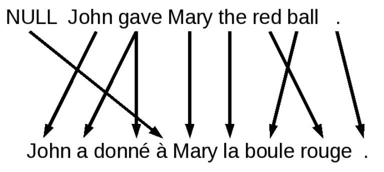
The SMT Problem



The IBM Learning Formulation - Brown et al. (1993)

- Find most likely word alignments
- Use alignments to generate translation table

Statistical Machine Translation
DETOUR



The IBM Assumptions

The source word generates the target word(s) based on:

- M1 - Lexical: identities of source and target words
- M2 - Distortion: location of source and target in their respective sentences
- M3 - Fertility: likelihood of source word to generate multiple targets
- Null: account for "spontaneous generated" targets

END DETOUR Convergence END DETOUR

Similarities

- Detect relationships between words
- Take into account similar factors:

	IBM	DMV
Type Association	Lexical	PoS
Relative Location	Distortion	Dist/Dir
Many-to-One	Fertility	Valence
Sourceless Targets	Null	Root

- Modular & incremental framework

Similarities

- Detect relationships between words
- Take into account similar factors:

	IBM	DMV
Type Association	Lexical	PoS
Relative Location	Distortion	Dist/Dir
Many-to-One	Fertility	Valence
Sourceless Targets	Null	Root

- Modular & incremental framework

Gibbs sampling implementation of IBM models - Thanks Chris!

EM vs. Gibbs Sampling

- EM: Clever counting of all possible alignments

+ very fast
- restrictive

EM vs. Gibbs Sampling

- EM: Clever counting of all possible alignments

+ very fast
- restrictive

- Sampling: Small transitions between alignments

+ easy to extend and add models
+ easy to experiment
- much slower

EM vs. Gibbs - Example

- IBM Model 3 (Brown et al. 1993, Equation 32):

$$P(\mathbf{f}|\mathbf{e}) = \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \binom{m-\phi_0}{\phi_0} p_0^{l-2\phi_0} p_1^{\phi_0} \prod_{j=1}^{l} n(\phi_i|e_i) \times$$

$$\prod_{j=1}^{m} t(f_i|e_{a_i}) d(j|a_j, m, l)$$

- Gibbs transition probabilities:

$$P(A[i] = j \Rightarrow \hat{j}\ ) \sim$$

$$\frac{P(\hat{A})}{P(A)} = \frac{P(w_j, \#deps(j) - 1) P(w_{\hat{j}}, \#deps(\hat{j}) + 1)}{P(w_j, \#deps(j)) P(w_{\hat{j}}, deps(\hat{j}))}$$

Model Construction

- Ideally, build dedicated models

Model Construction

- Ideally, build dedicated models
- Practically, start with what we have

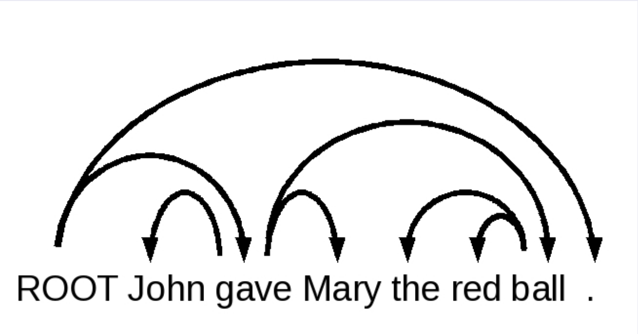
Model Construction

- Ideally, build dedicated models
- Practically, start with what we have
- IBM models address many necessary factors

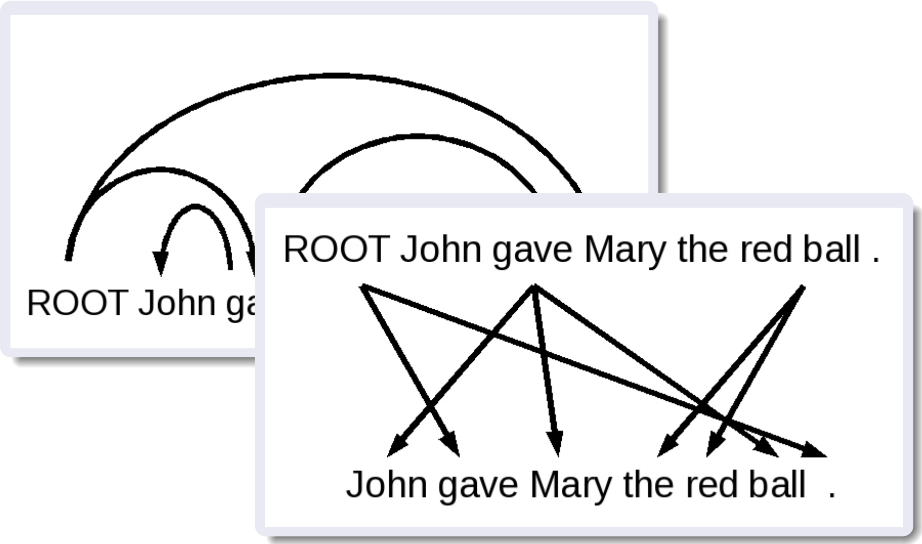
Model Construction

- Ideally, build dedicated models
- Practically, start with what we have
- IBM models address many necessary factors
- Experiment and improve as we go

Applying the Idea



Applying the Idea



Modeling Issues

Immediate Concerns

- Self alignments
- prevent words from choosing themselves as heads

Modeling Issues

Immediate Concerns

- Self alignments
  - prevent words from choosing themselves as heads
- Distortion model
  - distances (and direction) more relevant than location

Data

Datasets

- English - Penn. Treebank portion of the Wall Street Journal
- Danish and Dutch datasets from CoNLL 2006 shared task

Format

- Gold standard PoS tags
- Remove punctuation
- Sentences with ≤ 10

Results

Corpus	M 1	M2	M3	R-br
WSJ10	25.42	35.73	39.32	32.85
Dutch10	25.17	32.46	35.28	28.42
Danish10	23.12	25.96	41.94	16.05 *

- Model 2 beats baseline

Results

Corpus	M 1	M2	M3	R-br
WSJ10	25.42	35.73	39.32	32.85
Dutch10	25.17	32.46	35.28	28.42
Danish10	23.12	25.96	41.94	16.05 *

- Model 2 beats baseline
- Klein and Manning (2004): 43.2% for DMV

Results

Corpus	M 1	M2	M3	R-br
WSJ10	25.42	35.73	39.32	32.85
Dutch10	25.17	32.46	35.28	28.42
Danish10	23.12	25.96	41.94	16.05 *

- Model 2 beats baseline
- Klein and Manning (2004): 43.2% for DMV
- Surprisingly good for non-dedicated model!

Model 1 - Word Type Association

PoS	attachment
NN	DET
IN	NN
NNP	NNP
DET	NN
JJ	NN

PoS	attachment
NNS	JJ
RB	VBZ
VBD	NN
VB	TO
CC	NNS

Model 1 - Word Type Association

PoS	attachment
NN	DET
IN	NN
NNP	NNP
DET	NN
JJ	NN

PoS	attachment
NNS	JJ
RB	VBZ
VBD	NN
VB	TO
CC	NNS

- Detects dependency relations

Model 1 - Word Type Association

PoS	attachment
NN	DET
IN	NN
NNP	NNP
DET	NN
JJ	NN

PoS	attachment
NNS	JJ
RB	VBZ
VBD	NN
VB	TO
CC	NNS

- Detects dependency relations
- Directionality is a problem

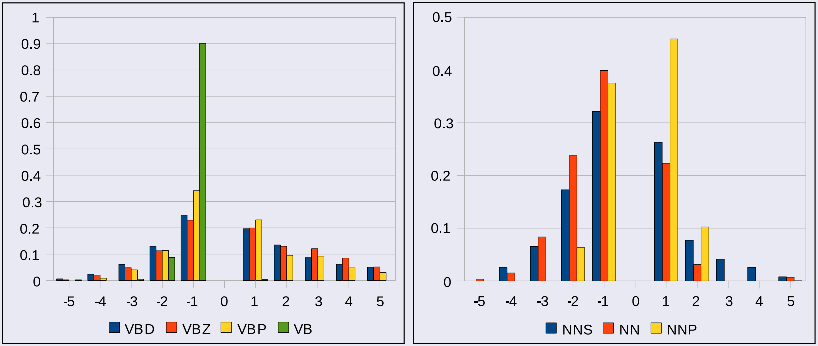
Model 1 - Word Type Association

PoS	attachment
NN	DET
IN	NN
NNP	NNP
DET	NN
JJ	NN

PoS	attachment
NNS	JJ
RB	VBZ
VBD	NN
VB	TO
CC	NNS

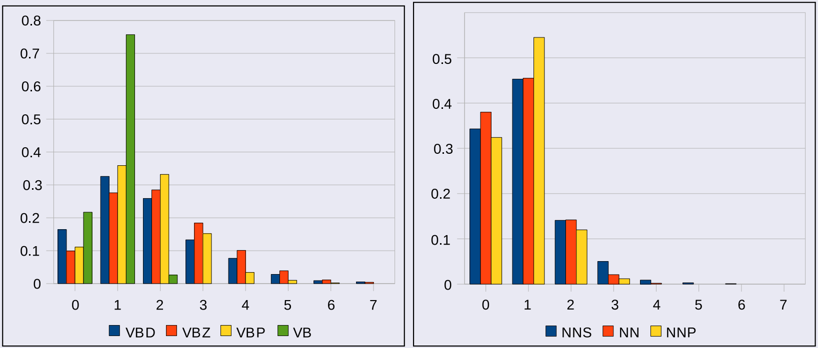
- Detects dependency relations
- Directionality is a problem
- **Note to self:** prevent cycles!

Model 2 - Distance



- Verbs have wider attachments
- Infinitives (VB) attach one to the left (TO)
- Proper nouns attach forward (no DET)

Model 3 - Fertility



- Verbs have wider fertility distribution
- Infinitives mostly have a single dependent

Conclusions

Lessons

- IBM models are a good start
- Minor adjustments necessary
- Further adjustments beneficial

Conclusions

Lessons

- IBM models are a good start
- Minor adjustments necessary
- Further adjustments beneficial

Framework

- Easy to extend and combine
- Easy to evaluate components
- Modular, sampling-based approach works

Future Work

Improving the Model

- Enforce tree structure
- Separate left and right
- Use lexical information
- Model root node better

Future Work

Improving the Model

- Enforce tree structure
- Separate left and right
- Use lexical information
- Model root node better

Extensions and Applications

- Incremental Learning (following Spitkovsky et al. 2010)
- Dependency-based SMT (e.g., Burkett et al. 2010)

Thank You!

Bibliography

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19(2):263–311.

Burkett, David, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *North American Association for Computational Linguistics*. Los Angeles.

Carroll, Glenn and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. In *Working Notes of the Workshop Statistically-Based NLP Techniques*. AAAI, pages 1–13.

Cohen, Shay B. and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, pages 74–82.

Headden III, William P., Mark Johnson, and David McClosky. 2009. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, pages 101–109.

Klein, Dan and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, page 478.

Paskin, Mark A. 2001. Grammatical bigrams. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*. MIT Press, pages 91–97.

Spitkovsky, Valentin I., Hiyan Alshawi, and Daniel Jurafsky. 2010. From Baby Steps to Leapfrog: How "Less is More" in unsupervised dependency parsing. In *Proc. of NAACL-HLT*.

Yuret, D. 1998. *Discovery of linguistic relations using lexical attraction*. Ph.D. thesis, Department of Computer Science and Electrical Engineering, MIT.