Chinese Natural Language Processing and Speech Processing

Overview

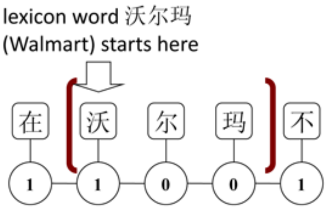
We work on a wide variety of research in Chinese Natural Language Processing and speech processing, including word segmentation, part-of-speech tagging, syntactic and semantic parsing, machine translation, disfluency detection, prosody, and other areas. We provide softwares for Chinese word segmentation, Chinese parsing and Chinese part-of-speech tagging.

More details on each topic:

- Chinese Word Segmentation
- Parsing and Grammatical Relations
- Part-of-Speech Tagging
- Named Entity Recognition
- Speech Processing

Chinese Word Segmentation

Our Chinese word segmenter relies on a linear-chain conditional random filed (CRF) model, which treats word segmentation as a binary decision task. It uses three categories of features: character identity n-grams, morphological and character reduplication features. As shown in the figure on the right, it also exploits lexicons and proper noun features to improve segmentation consistency, which is beneficial in tasks such as machine translation (MT) and information retrieval. In (Chang et al., 2008), we show that this increase of consistency yields an improvement of 0.32 BLEU point on a standard test set, and an additional improvement of 0.73 BLEU when segmentation granularity is optimized for the MT task.

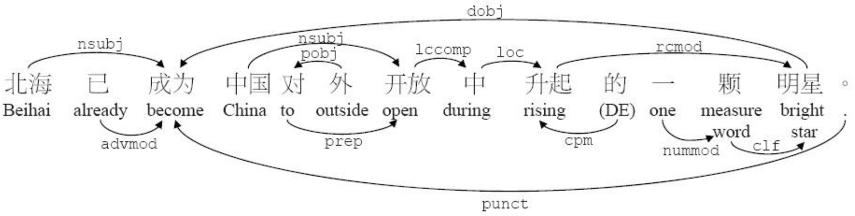


Parsing and Grammatical Relations

The Chinese parser is based on the ACL 2003 paper:

> Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? (http://nlp.stanford.edu/~manning/papers/acl2003-chinese.pdf). *ACL 2003*.

In addition to PCFG parsing, the Stanford Chinese parser can also output a set of Chinese grammatical relations that describes more semantically abstract relations between words. An example Chinese sentence looks like:


dobj
nsubj
nsubj
pobj
lccomp
loc
rcmod
北海 已 成为 中国 对 外 开放 中 升起 的 一 颗 明星 。
Beihai already become China to outside open during rising (DE) one measure word bright star .
advmod
prep
cpm
nummod
clf
punct

Details of the Chinese grammatical relations are in the 2009 SSST paper:

> Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features (http://nlp.stanford.edu/pubs/ssst09-chang.pdf).

Part-of-Speech Tagging

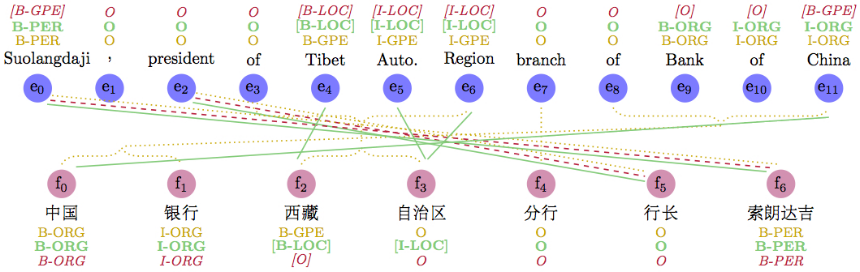
The Stanford part-of-speech tagger takes word-segmented Chinese text as input and assigns a part of speech to each word (and other tokens), such as a noun or a verb. This Chinese POS tagger is designed for LDC style word segmented texts, and adopts a subset of features from:

> Huihsin Tseng, Daniel Jurafsky, Christopher Manning. 2005. Morphological features help POS tagging of unknown words across language varieties (http://www.stanford.edu/~jurafsky/sighan_pos.pdf).

Its overall accuracy is 93.65% and the unknown word accuracy is 84.84%.

Named Entity Recognition

We have done extensive research on improving Chinese NER performance using semi-supervised learning methods with bilingual parallel text. Our results yield significant (~3% F1) improvements over strong CRF baselines that are enhanced with distributional similarity features.



[B-GPE] B-PER B-PER Suolangdaji e0
O O O , e1
O O O president e2
O O O of e3
[B-LOC] [B-LOC] B-GPE Tibet e4
[I-LOC] [I-LOC] I-GPE Auto. e5
[I-LOC] [I-LOC] I-GPE Region e6
O O O branch e7
O O O of e8
[O] B-ORG B-ORG Bank e9
[O] I-ORG I-ORG of e10
[B-GPE] I-ORG I-ORG China e11
f0 中国 B-ORG B-ORG B-ORG
f1 银行 I-ORG I-ORG I-ORG
f2 西藏 B-GPE [B-LOC] [O]
f3 自治区 O [I-LOC] O
f4 分行 O O O
f5 行长 O O O
f6 索朗达吉 B-PER B-PER B-PER

Details of the semi-supervised learning with bitext work can be found in the following papers:

> Mengqiu Wang and Christopher D. Manning. 2013. Cross-lingual Pseudo-Projected Expectation Regularization for Weakly Supervised Learning (http://cs.stanford.edu/people/mengqiu/publication/tacl13.pdf). *Transactions of ACL 2013*
> Mengqiu Wang, Wanxiang Che and Christopher D. Manning. 2013. Joint Word Alignment and Bilingual Named Entity Recognition Using Dual Decomposition (http://cs.stanford.edu/people/mengqiu/publication/acl13.pdf). *ACL 2013*
> Mengqiu Wang, Wanxiang Che and Christopher D. Manning. 2013. Effective Bilingual Constraints for Semi-supervised Learning of Named Entity Recognizers (http://cs.stanford.edu/people/mengqiu/publication/aaai13.pdf). *AAAI 2013* (Outstanding Paper Award Honorable Mention)
> Wanxiang Che, Mengqiu Wang and Christopher D. Manning. 2013. Named Entity Recognition with Bilingual Constraints (http://cs.stanford.edu/people/mengqiu/publication/naacl13.pdf). *NAACL 2013*

Software Instructions:

> Follow these instructions (biNER-instruction.shtml) to reproduce experiments reported in these papers.

Speech Processing

Our Chinese speech research has focused on areas like the study and detection of disfluencies (filled pauses like *uh* and word fragments), prosody, and the detection of speech acts.

People

- Members:
  - Dan Jurafsky (http://www.stanford.edu/~jurafsky/)
  - Christopher Manning (http://nlp.stanford.edu/~manning/)
- Alumni/Alumnae:
  - Galen Andrew (https://www.cs.washington.edu/homes/galen/), now a graduate student at the University of Washington
  - Pi-Chuan Chang (http://nlp.stanford.edu/~pichuan/), now at Google Research
  - Michel Galley (http://www-nlp.stanford.edu/~mgalley/), now at Microsoft Research
  - Roger Levy (http://idiom.ucsd.edu/~rlevy/), now at UCSD
  - Huihsin Tseng, now at Yahoo!
  - Yun-Hsuan Sung (http://www.stanford.edu/~yhsung/), now at Google Research

Software

- The Stanford Chinese Segmenter (http://nlp.stanford.edu/software/segmenter.shtml)
- The Stanford Chinese Part-of-Speech Tagger (http://nlp.stanford.edu/software/tagger.shtml)
- The Stanford Chinese Named Entity Recognizer (http://nlp.stanford.edu/software/CRF-NER.shtml)
- The Stanford Chinese Parser (http://nlp.stanford.edu/software/lex-parser.shtml)

Publications

Cross-lingual Pseudo-Projected Expectation Regularization for Weakly Supervised Learning [pdf (http://cs.stanford.edu/people/mengqiu/publication/tacl13.pdf)]
Mengqiu Wang and Christopher D. Manning.
in Transactions of ACL, 2013.

Joint Word Alignment and Bilingual Named Entity Recognition Using Dual Decomposition [pdf (http://cs.stanford.edu/people/mengqiu/publication/acl13.pdf)]
Mengqiu Wang, Wanxiang Che and Christopher D. Manning.
in Proceedings of ACL, 2013.

Effective Bilingual Constraints for Semi-supervised Learning of Named Entity Recognizers [pdf (http://cs.stanford.edu/people/mengqiu/publication/aaai13.pdf)]
Mengqiu Wang, Wanxiang Che and Christopher D. Manning.
in Proceedings of AAAI, 2013.

Named Entity Recognition with Bilingual Constraints [pdf (http://cs.stanford.edu/people/mengqiu/publication/naacl13.pdf)]
Wanxiang Che, Mengqiu Wang and Christopher D. Manning.
in Proceedings of NAACL, 2013.

Discriminative Reordering with Chinese Grammatical Relations Features [pdf (doc/ssst09-cameraready.pdf)]
Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning.
in NAACL 2009 Third Workshop on Syntax and Structure in Statistical Translation.

Disambiguating "DE" for Chinese-English Machine Translation [pdf (doc/wmt09.pdf)]
Pi-Chuan Chang, Dan Jurafsky and Christopher D. Manning.
in EACL 2009 Fourth Workshop on Statistical Machine Translation.

Optimizing Chinese Word Segmentation for Machine Translation Performance [pdf (doc/acl-wmt08-cws.pdf)]
Pi-Chuan Chang, Michel Galley and Christopher D. Manning.
in ACL 2008 Third Workshop on Statistical Machine Translation.

Stanford University's Chinese-to-English Statistical Machine Translation System for the 2008 NIST Evaluation [pdf (doc/nist08.pdf)]
Michel Galley, Pi-Chuan Chang, Daniel Cer, Jenny R. Finkel, Christopher D. Manning.
in Proceedings of the 2008 NIST Open Machine Translation Evaluation Workshop.

Detection of Word Fragments in Mandarin Telephone Conversation [pdf (icslp06fragfinal.pdf)]
Cheng-Tao Chu, Yun-Hsuan Sung, Yuan Zhao, Dan Jurafsky.
Proceedings of INTERSPEECH-2006, Pittsburgh, PA.

A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005 [pdf (http://nlp.stanford.edu/pubs/sighan2005.pdf)]
Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning
The Fourth SIGHAN Workshop on Chinese Language Processing, 2005

Morphological features help POS tagging of unknown words across language varieties [pdf (http://www.stanford.edu/~jurafsky/sighan_pos.pdf)]
Huihsin Tseng, Daniel Jurafsky, Christopher Manning
The Fourth SIGHAN Workshop on Chinese Language Processing, 2005

Accent Detection and Speech Recognition for Shanghai-Accented Mandarin [pdf (p1304.pdf)]
Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Dan Jurafsky, Rebecca Starr and Su-Youn Yoon.
Proceedings of EUROSPEECH-05

A preliminary study of Mandarin filled pauses [pdf (diss05.pdf)]
Yuan Zhao and Dan Jurafsky
Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop

Detection of Questions in Chinese Conversation [pdf (yuan8997.pdf)]
Yuan, Jiahong and Dan Jurafsky
Proceedings of IEEE ASRU 2005

Parsing Arguments of Nominalizations in English and Chinese [pdf (hlt-2004-noun.pdf)] Pradhan, Sameer, Honglin Sun, Wayne Ward, James H. Martin, and Daniel Jurafsky
Proceedings of NAACL-HLT 2004.

Is it harder to parse Chinese, or the Chinese Treebank? [pdf (http://nlp.stanford.edu/pubs/acl2003-chinese.pdf)]
Roger Levy and Christopher Manning
Proceedings of ACL 2003

Stanford NLP Group

Gates Computer Science Building
353 Serra Mall
Stanford, CA 94305-9010
Directions and Parking (http://forum.stanford.edu/visitors/directions/gates.php)

Affiliated Groups

- **Stanford AI Lab (http://ai.stanford.edu/)**
- **Stanford InfoLab (http://infolab.stanford.edu/)**
- **Center for the Study of Language and Information (https://www-csli.stanford.edu/)**

Connect

- Stack Overflow (http://stackoverflow.com/tags/stanford-nlp)
- Github (https://github.com/stanfordnlp/CoreNLP)
- Twitter (https://twitter.com/stanfordnlp)

Local links: NLP lunch (/local/nlp_lunch.shtml) · NLP Reading Group (http://nlp.stanford.edu/read/) · NLP Seminar (http://nlp.stanford.edu/seminar/) · AI Speakers (http://ai.stanford.edu/portfolio-view/distinguished-speaker-series) · JavaNLP (/javanlp/) (javadocs (/nlp/javadoc/javanlp/)) · machines (/local/machines.shtml) · Wiki (/wiki/) · Calendar (/local/calendar.shtml) · Q&A (/local/qa/)