One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction

Hwee Tou Ng
Department of Computer Science
National University of Singapore

8 Jan 2018

Joint work with Kaveh Taghipour

Introduction

- Goal: Automatically identify the meaning (sense) of a word based on its context

- Two main kinds:
  - Word Sense Disambiguation (WSD)
    - A classification task
    - Based on a predefined set of senses in an existing sense inventory (e.g., WordNet)
  - Word Sense Induction (WSI)
    - A clustering task
    - Group each instance of a target ambiguous word with some other instances to form a cluster of instances

Motivation

- Supervised machine learning approach gives the best performance for word sense disambiguation (WSD)
- Drawback: These systems need annotated training data
- Bottleneck: Lack of large-scale manually annotated sense-tagged data
- Very few large annotated datasets are available to the research community

Resource Building

- WordNet: A valuable resource for specifying word senses (meanings of words) in English

- Sense-tagged corpora: An additional resource needed in large quantities for building automatic word sense disambiguation systems

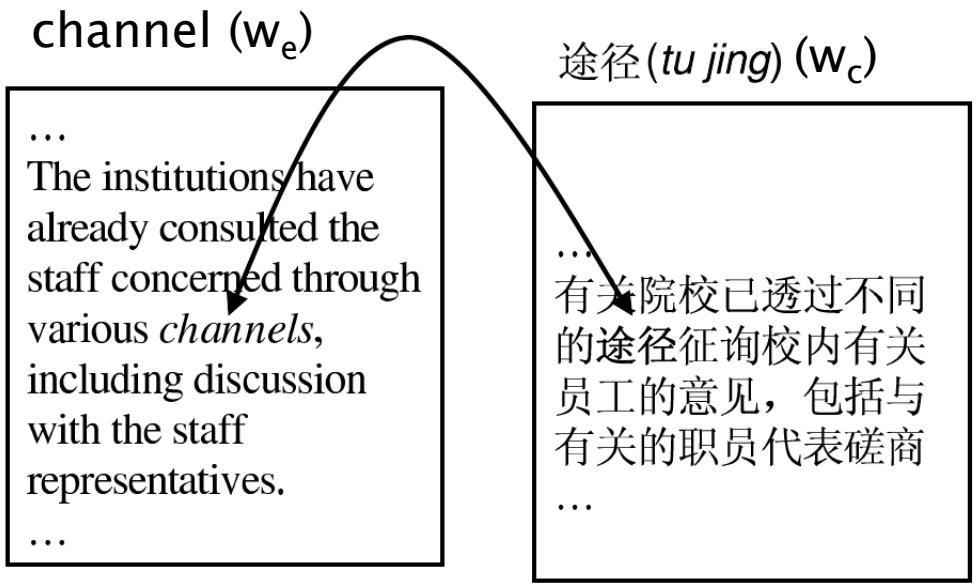
Objectives

- Extract and annotate a large number of sense-tagged instances and make them publicly available for research purpose
- Evaluate the use of this dataset on word sense disambiguation and induction tasks
- Work published in (Taghipour & Ng, CoNLL 2015)

Base Corpus

- MultiUN parallel corpus (MUN):
  - A collection of translated documents from the United Nations (Eisele & Chen, LREC 2010)
  - Six official languages of the UN (Arabic, Chinese, English, French, Russian, Spanish)
  - A freely available parallel corpus
  - Produced in the EuroMatrixPlus project
  - An automatically sentence-aligned version of this dataset can be downloaded from the OPUS website
  - We use the Chinese-English portion of MultiUN parallel corpus

Sense-Tagged Data via Parallel Texts



ID	Description	Translations
1	A path over which electrical signals can pass	频道
2	A passage for water	水道, 水渠, 排水渠
3	A long narrow furrow	沟
4	A relatively narrow body of water	海峡
5	A means of communication or access	途径
6	A bodily passage or tube	导管
7	A television station and its programs	频道

Sense-Tagged Data via Parallel Texts

- To assign a sense tag to an English word w_e (e.g., channel) in a sentence, we make use of the aligned Chinese translation w_c (e.g., 途径) of w_e based on automatic word alignment.
- For each sense *i* of w_e in the sense inventory, a list of Chinese translations of sense *i* of w_e has been manually created.
- If w_c matches one of these Chinese translations of sense *i*, then w_e is tagged with sense *i*.

Preprocessing of Parallel Texts

- Tokenization (English part)
- Word segmentation (Chinese part)
- Word alignment (via GIZA++)
- Part-of-speech (POS) tagging and lemmatization of English words

Coverage of Annotated Corpus

- Number of word types covered:
  - 649 nouns
  - 190 verbs
  - 319 adjectives

- Encompasses the top 60% most frequent English word types (for nouns, verbs, adjectives) based on frequency in the Brown corpus

Quality of Annotated Corpus

- Manually evaluated 1,000 randomly selected sense-tagged instances
- Sense-tag accuracy: 83.7% (based on the fine-grained sense inventory of WordNet 1.7.1)
- Error analysis
  - 4% of errors due to wrong sentence or word alignment
  - 69% of errors due to one Chinese word being the translation of multiple English senses

Annotated Corpus

- To speed up the training process, we perform random sampling on senses with more than 500 instances and limit the number of selected instances per sense to 500
- However, all senses with fewer than 500 instances are included in the training data
- This sampling method ensures that rare sense tags also have training instances

Annotated Corpus

- We augment the dataset by adding sense-tagged instances from SEMCOR (Miller et al, HLT 1993) and the DSO corpus (Ng & Lee, ACL 1996)

- We convert the sense tags using the sense mapping files provided by WordNet and release our sense-tagged corpus in three WordNet versions (1.7.1, 2.1, 3.0)

Statistics of Annotated Corpus

- Number of word types in each part-of-speech

	noun	verb	adjective	adverb	total
MUN (before sampling)	649	190	319	0	1,158
MUN	649	190	319	0	1,158
MUN+SC	11,446	4,705	5,129	28	21,308
MUN+SC+DSO	11,446	4,705	5,129	28	21,308

Statistics of Annotated Corpus

- Number of training instances in each part-of-speech

	number of training samples				
	noun	**verb**	**adjective**	**adverb**	**total**
MUN (before sampling)	14,492,639	4,400,813	4,078,543	0	22,971,995
MUN	503,408	265,785	218,046	0	987,239
MUN+SC	582,028	341,141	251,362	6,207	1,180,738
MUN+SC+DSO	687,871	412,482	251,362	6,207	1,357,922

Statistics of Annotated Corpus

- Average number of instances per word type

	Avg. # samples per word type
MUN (before sampling)	19,837.6
MUN	852.5
MUN+SC	55.4
MUN+SC+DSO	63.7

Evaluation

- Supervised WSD system used:
  - IMS (It Makes Sense) (Zhong & Ng, ACL 2010)
  - Widely used in the WSD research community for comparison on benchmark test data
- Based on support vector machines (SVM)
- Features:
  - POS tags
  - Collocations
  - Surrounding words

Evaluation

- All-words word sense disambiguation tasks:
  - SensEval-2 (fine-grained)
  - SensEval-3 task 1 (fine-grained)
  - SemEval-2007 task 17 (fine-grained)
  - SemEval-2007 task 7 (coarse-grained)

- Word sense induction task:
  - SemEval-2013 task 13

Evaluation

- Accuracy (in %) on all-words WSD tasks

	SensEval-2	SensEval-3	SemEval-2007	
	Fine	Fine	Fine	Coarse
IMS (MUN)	64.5	60.6	52.7	78.7
IMS (MUN+SC)	68.2	67.4	58.5	81.6
IMS (MUN+SC+DSO)	68.0	66.6	58.9	82.3
IMS (original)	68.2	**67.6**	58.3	**82.6**
IMS (SC)	66.1	67.0	58.7	81.9
IMS (SC+DSO)	66.5	67.0	57.8	81.6
Rank 1	**69.0**	65.2	**59.1**	82.5
Rank 2	63.6	64.6	58.7	81.6
WordNet Sense 1	61.9	62.4	51.4	78.9

Evaluation

- Evaluation results (in %) on WSI task

	Jac. Ind.	K_{δ}^{sim}	**WNDCG**	**F. NMI**	**F. B-Cubed**
IMS (MUN)	24.6	64.9	33.0	6.9	57.1
IMS (MUN+SC)	25.0	**65.4**	34.2	9.1	55.9
IMS (MUN+SC+DSO)	**25.5**	**65.4**	35.1	**9.7**	55.4
IMS (original)	23.4	64.5	34.0	8.6	59.0
IMS (SC)	22.9	63.5	32.4	6.8	57.3
IMS (SC+DSO)	23.4	63.6	32.9	7.1	57.6
Wang-15 (ukWac)	-	-	-	**9.7**	54.5
Wang-15 (actual)	-	-	-	9.4	59.1
AI-KU (base)	19.7	62.0	**38.7**	6.5	39.0
AI-KU (add1000)	19.7	60.6	21.5	3.5	32.0
AI-KU (remove5-add1000)	24.4	64.2	33.2	3.9	45.1
Unimelb (5p)	21.8	61.4	36.5	5.6	45.9
Unimelb (50k)	21.3	62.0	37.1	6.0	48.3
all-instances-1cluster	19.2	60.9	28.8	0.0	**62.3**
each-instance-1cluster	0.0	0.0	0.0	7.1	0.0
SEMCOR most freq sense	19.2	60.9	28.8	0.0	**62.3**

Conclusions

- The major problem in building supervised word sense disambiguation systems is the training data acquisition bottleneck

- We semi-automatically extracted and sense-tagged an English corpus containing one million sense-tagged instances

Conclusions

- Our sense-tagged corpus
  - Publicly released since 2015:
    (http://www.comp.nus.edu.sg/~nlp/corpora.html)
  - Used to build a WSD system that performs competitively with the top performing WSD systems in several all-words WSD tasks, and the top systems in a WSI task
  - Used in subsequent work by other researchers