Lexical Disambiguation

The Interaction of Knowledge Sources in Word Sense Disambiguation

Will Roberts
wroberts@coli.uni-sb.de

Wednesday, 4 June, 2008

Introduction

- Little consensus on the correct way to do Word Sense Disambiguation
- Choices:
  - limited vocabulary or broad-coverage?
  - supervised or unsupervised?
  - granularity: sense or homograph level?
- Syntactic, semantic and pragmatic information can all be useful sources of information for WSD:
  1. John did not feel *well*.
  2. John tripped near the *well*.
  3. The *bat* slept.
  4. He bought a *bat* from the sports shop.

Multiple Knowledge Sources

Ng and Lee (1996) tagged word senses for the word *interest* in the *Wall Street Journal* using a k-nearest neighbor learning algorithm:

Table 1
Relative contribution of knowledge sources in LEXAS.

Knowledge Source	Accuracy
Collocations	80.2%
PoS and Morphology	77.2%
Surrounding words	62.0%
Verb-object	43.5%

Lexicon

Longman Dicionary of Contemporary English:

- designed for students of English
- 36,000 word types, with senses grouped into homographs
- words with one closely grouped set of senses are *monohomographic*

Word Senses

bank^{1} *n* **1** land along the side of a river, lake, etc. **2** earth which is heaped up in a field or a garden, often making a border or division **3** a mass of snow, mud, clouds, etc.: *The banks of dark cloud promised a heavy storm* **4** a slope made at bends in a road or race-track, so that they are safer for cars to go round **5** SANDBANK: *The Dogger Bank in the North Sea can be dangerous for ships*

bank^{2} *v* [IØ] (of a car or aircraft) to move with one side higher than the other, esp. when making a turn – see also BANK UP

bank^{3} *n* **1** a row, esp. of OARs in an ancient boat or KEYs on a TYPEWRITER

bank^{4} *n* **1** a place where money is kept and paid out on demand, and where related activities go on – see picture at STREET **2** (*usu. in comb.*) a place where something is held ready for use, esp. ORGANIC product of human origin for medical use: *Hospital bloodbanks have saved many lives* **3** (a person who keeps) a supply of money or pieces for payment or use in a game of chance **4 break the bank** to win all the money that the BANK^{4}(3) has in a game of chance

bank^{5} *v* **1**[T1] to put or keep (money) in a bank **2**[L9, esp. *with*] to keep one's money (esp. in the stated bank): *Where do you bank?*

Homographs

- each homograph is marked with a part of speech
- about 2% of words have a homograph with more than one part of speech (usually noun and verb)
- homograph groupings are fairly course, however this is often sufficient (e.g., for translation equivalents):
    - "financial institution" translates to *banque* in French;
    - "edge of river" is *bord*

Disambiguation using Part of Speech

- 34% of content words in LDOCE are polysemous, but only 12% are polyhomographic
- Thus, part of speech can disambiguate 88% of words to the homograph level
- Some words can be disambiguated to this level if they have certain part of speech tags, but not others:
  - *beam* has 3 homographs: 2 which are nouns and 1 which is a verb
- 7% of words are of this type
- Theoretically, 95% of words could be disambiguated to the homograph level by part of speech alone

Quantifying the Part of Speech Contribution

- Five articles from *Wall Street Journal* containing 391 polyhomographic words
- Correct homograph senses were manually annotated by authors for a gold standard
- The texts were then tagged using a Brill tagger
- If a word had more than one homograph with the same POS, the most frequently occurring sense was chosen
- 87.4% of polyhomographic words were assigned the correct homograph
- Baseline: choose the most frequent homograph regardless of POS information
  - ⇒ 78% of tokens were correctly disambiguated this way

Part of Speech Filtering

The POS tagger is run over the text, and homographs with non-matching POS are removed.

- Full disambiguation: only a single homograph remains
- Partial disambiguation: several homographs remain, but some have been removed from consideration
- No disambiguation: all the homographs of a word have the same POS
- POS error: the correct homograph is removed from consideration through tagger error. Sometimes all possible homographs are filtered out by these kinds of errors.

Part of Speech Filtering

Table 3
Error analysis for the experiment on WSD by part of speech alone.

Word Type	Count	Correctly disambiguated by: Baseline method	PoS method
Full disambiguation	297	268 (90%)	297 (100%)
Partial disambiguation	58	22 (38%)	32 (55%)
No disambiguation	23	10 (43%)	10 (43%)
Part-of-speech error	13	5 (38%)	3 (23%)
All polyhomographic	391	305 (78%)	342 (87%)

Part of Speech Filtering

Table 2
Examples of the four word types introduced in Section 3.2. The leftmost column indicates the full set of homographs for the example words, with upper case indicating the correct homograph. The remaining columns show (respectively) the part-of-speech assigned by the tagger, the resulting set of senses after filtering, and the type of the word.

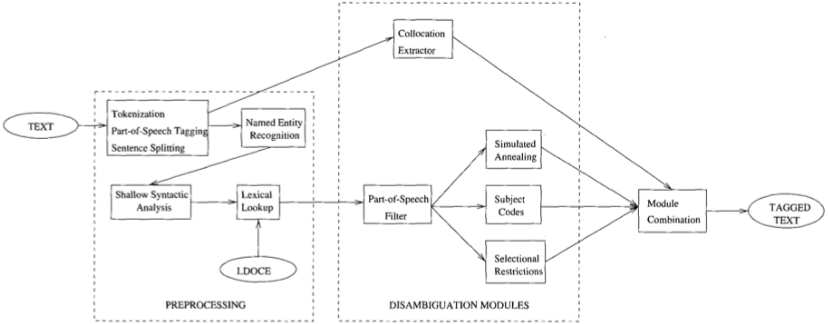
All Homographs	PoS Tag	After tagging	Word type
N, v, v	n	N	Full disambiguation
n, adj, V	v	V	Full disambiguation
n, V, v	v	V, v	Partial disambiguation
n, N, v	n	n, N	Partial disambiguation
N, n	n	N, n	No disambiguation
v, V	v	v, V	No disambiguation
N, v, v	v	v v	PoS error
N, v, v	adj	N, v, v	PoS error

Framework for Combining Knowledge Sources

Modular architecture composed of:

- filters: remove senses from consideration when they appear to be unlikely in context
- partial taggers: representing evidence for or against a particular sense, but with lower confidence
- feature extractors: representing the context of ambiguous words

Framework for Combining Knowledge Sources



Preprocessing

Initial stage of framework.

1. tokenization
2. lemmatization
3. split into sentences
4. POS tagging, using the Brill tagger
5. Named Entity Recognition

Preprocessing

Scope of disambiguation after preprocessing:

- only content words (can be identified by part of speech tag)
- no disambiguation of words inside named entities (since they are usually analyzed by the named entity identifier)

Partial Tagger: Simulated Annealing

Based on measuring the overlap of dictionary definitions, e.g., *bank* and *river*.

- Measuring the dictionary definition overlap in this way for every possible combination of senses for every word in a sentence is too computationally demanding.
- Solution is approximated using simulated annealing.
- Cowie, Guthrie, and Guthrie (1992), using LDOCE, found this could disambiguate 47% of words to the sense level, and 72% to the homograph level, compared to manually assigned senses.
- Distance metric used is a normalized count of the number of words overlapping between two definitions.

Partial Tagger: Selectional Preferences

Based on finding the set of senses for each word that are licensed by selectional preferences.

- LDOCE senses are marked with selectional restrictions indicated by 36 semantic codes.
- These are arranged into a hierarchy to deal with varying levels of generality.
- named entities identified in preprocessing can also be used by this module

Partial Tagger: Selectional Preferences

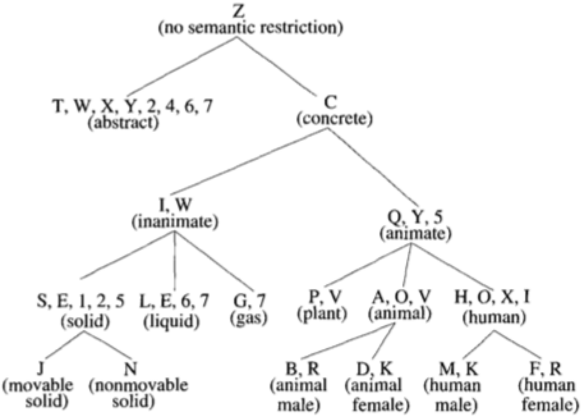


Figure 3
Bruce and Guthrie's hierarchy of LDOCE semantic codes.

Partial Tagger: Selectional Preferences

Sense selection starts at the verb and extends to the verb's dependencies, etc.

1. Syntactic relationships in the sentence are identified by a shallow parser, which finds subject-verb, direct object, indirect object and noun-adjective relations.
   - The parser has achieved 51% precision and 69% recall when tested against the Penn Tree Bank.
2. Each sense of a verb applies a preference to the subject and object nouns, which may disallow some senses for these.
   - If a sense of a verb disallows all senses of one of its dependent nouns, that verb sense is immediately rejected.
3. For each noun that is modified by an adjective, we can again filter the adjective senses that do not agree with any of the remaining noun senses.

Partial Tagger: Selectional Preferences

Table 5
Sentence and lexicon for toy example of selectional preference resolution algorithm.

Example sentence:
John ran the hilly course.

Sense	Definition and *Example*	Restriction
John	proper name	type:human
ran (1)	to control an organisation *run IBM*	subject:human object:abstract
ran (2)	to move quickly by foot *run a marathon*	subject:human object:inanimate
hilly (1)	undulating terrain *hilly road*	modifies:nonmovable solid
course (1)	route *race course*	type:nonmovable solid
course (2)	programme of study *physics course*	type:abstract

Partial Tagger: Selectional Preferences

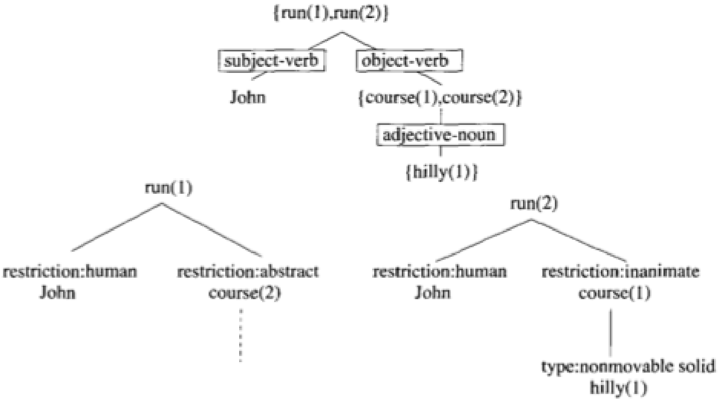


Figure 4
Restriction resolution in toy example.

Partial Tagger: Subject Codes

Based on categorization of word senses into subject areas; e.g., “Linguistics and Grammar” is assigned to some senses of the words “ellipsis”, “ablative”, “bilingual”, and “intransitive”.

- 56% of words in LDOCE have no subject code, and are assigned the code --.

$$\arg\max_{SCat} \sum_{w \in context} \log \frac{P(w|SCat)P(SCat)}{P(w)}$$

Partial Tagger: Subject Codes

- Prior probability $P(SCat)$ is estimated from the proportion of word senses in LDOCE assigned this subject code.
- Context of 50 words on either side of the ambiguous word is used.
- Word probabilities were collected from British National Corpus (14 million words), with no smoothing applied; only context words which appeared at least 10 times in the training data were used.
- Yarowsky (1992) reports 92% correct disambiguation on 12 test words with an average of 3 possible subject categories using Roget's thesaurus; however, LDOCE has higher ambiguity and a smaller thesaural hierarchy.

Collocation Extractor

10 collocates are extracted for each ambiguous word:

- first word to the left, first word to the right, second word to the left, second word to the right, first noun to the left, first noun to the right, first verb to the left, first verb to the right, first adjective to the left, first adjective to the right.
- Collocates are extracted from the current sentence; if a collocate does not exist, it is coded as `NoColl`.
- Morphological roots are stored instead of surface forms; this might help with data sparseness.

Combining Results

Results from the disambiguation modules are presented to a *k*-nearest neighbor algorithm called TiMBL.

This approach relies on a weighted distance metric:

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i)$$

$$\delta(x_i, y_i) = \begin{cases} \frac{x_i - y_i}{max_i - min_i} & \text{if numeric, else} \\ 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases}$$

Combining Results

Weights for each feature are based on a Gain Ration measure, which indicates the difference in uncertainty between the situations with and without knowledge of that feature:

$$w_i = \frac{H(C) - \sum_v P(v) \times H(C|v)}{H(v)}$$

C is the set of class labels, v ranges over all values of the feature i and H is entropy. The weighting is normalized by the entropy of the feature values, to cancel the effect of a feature with many possible values.

Combining Results

Context
Regarding Atlanta's new million dollar airport, the jury recommended "that when the new management take charge Jan. 1 the airport be operated in a manner that will eliminate political **influences**".

Feature Vectors

Learning features	Truth
influence 1 1a 1 n influences 1 12.03 y NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	correct
influence 1 1b 2 n influences 0 12.03 y NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	incorrect
influence 1 2 3 n influences 0 12.03 y NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	incorrect
influence 1 3 4 n influences 0 12.03 y NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	incorrect
influence 1 4 5 n influences 0 12.03 n NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	incorrect
influence 1 5 6 n influences 0 12.03 n NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	incorrect
influence 1 6 7 n influences 0 12.03 n NoColl manner NoColl eliminate NoColl in NoColl political NoColl eliminate	incorrect

Figure 5
Example feature-vector representation.

Evaluation

- Most strategies rely on a human-generated gold standard.
- This may be difficult for humans to do, and generating gold standards is very labor-intensive compared to POS tagging.
- Evaluation here combined two existing resources:
  - SEMCOR: part of the WordNet project, a 200,000 word corpus with the content words manually tagged
  - SENSUS: large-scale ontology designed for machine-translation, a merger of the ontologies of WordNet, LDOCE and the Penman Upper Model
- Evaluated on the collected data using 10-fold cross validation
- Exact match metric: ratio of correctly assigned senses to number of senses assigned

Evaluation

Zipfian distribution of ambiguous words:

Table 6
Occurrence of ambiguous words in the evaluation corpus.

Occurrence Range	Count
1–25	5488 (94.6%)
26–50	202 (3.5%)
51–75	67 (1.2%)
76–100	21 (0.04%)
100–604	26 (0.4%)

Evaluation

Table 7
System results, baselines, and corpus characteristics. Sense level results are calculated over all polysemous words in the evaluation corpus while those reported for the homograph level are calculated only over polyhomographic ones.

		Entire Corpus	Subcorpora Noun	Verb	Adjective	Adverb
Sense level	Accuracy	**90.37%**	91.24%	88.38%	91.09%	70.61%
	Baseline	**30.90%**	34.56%	18.46%	25.76%	36.73%
	Tokens	**36,774**	26,091	6,465	3,310	908
	Types	**5,804**	4.041	1,021	1,006	125
	Average Polysemy	**14.62**	13.65	24.35	6.07	4.43
Homograph level	Accuracy	**94.65%**	94.63%	95.26%	96.89%	90.67%
	Baseline	**71.24%**	73.47%	60.72%	87.10%	86.87%
	Tokens	**18,219**	11,380	5,194	1,326	319
	Types	**1,683**	1,264	709	201	34
	Average Polysemy	**2.52**	2.32	2.81	2.95	3.13

Performance of Individual Modules

Table 8
Performance of individual partial taggers (at sense level).

	All	Nouns	Verbs	Adjectives	Adverbs
simulated annealing (1)	65.24%	66.50%	67.51%	49.02%	50.61%
selectional preferences (2)	44.85%	40.73%	75.80%	27.56%	0%
subject codes (3)	79.41%	79.18%	72.75%	73.73%	85.50%

Conclusion

- Broad coverage word sense disambiguation system with high accuracy
- Uses a standard machine readable dictoinary
- More accurate results when many knowledge sources are combined
- Demonstrates the relative independence of the types of semantic information used
- Possible that WSD is a more difficult problem than part-of-speech, and that it may never achieve the precision of POS taggers.

Literature

Stevenson, M. and Wilks, Y. 2001.
The Interaction of Knowledge Sources in Word Sense Disambiguation.
Computational Linguistics, 27(3).