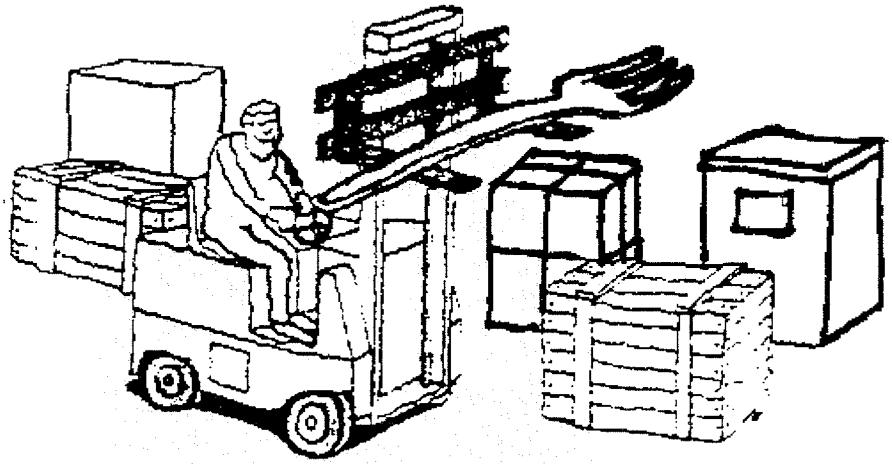
Pilot SENSEVAL

Herstmonceux Castle

September 1998



Hot-off-the-Press Papers

Under the auspices of:
ACL-SIGLEX
EURALEX
ELSNET
and EU Projects:
ECRAN
SPARKLE

SIGLEX Report - 1999 - Martha Palmer, Chair

The special issue of Natural Language Engineering that includes papers from the Siglex semantic tagging workshop at ANLP97 is in progress. The revised versions of the papers have been received and are undergoing a final review process. Editors Martha Palmer and Marc Light.

SENSEVAL and ROMANSEVAL, were held at Herstmonceux Castle in early September, 1998, organized by Adam Kilgarriff and Martha Palmer. There were 54 participants, 24 systems, 3 languages: 35 English words and 60 French and Italian words. Training data and test data was prepared for the English words, based on the Hector corpus, with approximately 200 corpus instances for each word for training purposes, and dozens of additional instances for testing purposes. The workshop participants were quite enthusiastic about the usefulness of this exercise, and pleased with the system performances. The human annotator agreement was over 90% while the systems approached Precision and recall figures in the low-80% range.

However, there was general agreement that the next evaluation should include tagged running text, and that the sense inventory being used should include sense distinctions with clear relevance to applications such as machine translation and information retrieval.

The proceedings will appear as a special issue of Computers and the Humanities.

http://www.itri.bton.ac.uk/events/senseval/cfp2.html,

and the acceptance notices for the papers have just gone out. Editors: Adam Kilgarriff and Martha Palmer.

ACL99 is the site for SIGLEX99, the 6th SIGLEX workshop, where in addition to papers we have working sessions for the discussion of samples of sense tagged running text. We are also discussing how WordNet could be revised to make it more suitable for sense tagging purposes, and will be planning our next Senseval around our conclusions, presumably Siglex2K. We are also continuing our discussions of American involvement in EAGLES, now know as ISLE, International Standards for Language Engineering. The new agenda for ISLE will be extending standards for lexical semantics based on American feedback, and including standards for linking entries in multilingual lexicons.

http://www.ilc.pi.cnr.it/EAGLES96/rep2/rep2.html

Finally, SIGLEX99 will be having a business meeting to discuss the election of officers and the adoption of a constitution.

1999-06-21 22:57 acl99.report

Programme Detail: Order of system demonstrations

7 mins max for each presentation plus max 5 mins questions: other questions to wait to the end.

Who	Research Group	Lg	System
Weds 2nd	2.30-4.00		
Chair	Frédérique Segond		
Diana McCarthy	Univ Sussex	Eng/o	sussex
AK for Ken Litkowski	CL Research	Eng/a	clres
Dekang Lin	Univ Manitoba	Eng/a	manitoba-dl
Ken Barker	Univ Ottawa	Eng/a	ottawa
Eneko Agirre	Tech Univ Catalonia, Univ Basque	Eng/a	upc-ehu-un
Jeremy Ellman	Univ Sunderland	Eng/a	suss
Romaric Besançon	EPFL	Fr	
Vito Pirelli	Pisa	It	

Thurs 3rd	9.50–11.00		
Chair	David Yarowsky		
Frédérique Segond	XRCE/CELI	Eng/a	xeroxceli
Frédérique Segond	XRCE	Fr	
Tom O'Hara	New Mex State, UNC Asheville	Eng/s	grling-sdm
Claudia Leacock	Educ Testing Service, Princeton	Eng/s	ets-pu
Paul Hawkins	Univ Durham	Eng/s	durham

Thurs 3rd	11.30–12.20		
Chair	Nicoletta Calzolari		
Hae-Chang Rim	KAIST, Korea	Eng/s	korea
David Yarowsky	John Hopkins Univ	Eng/s	hopkins
Keith Suderman	Univ Manitoba	Eng/s	manitoba-ks
Jorn Veestra	ITK, Tilburg	Eng/s	tilburg

Systems participating but unable to attend:
* clres (Ken Litkowski, Eng/a)
* avignon (Claude de Loupy, Bertin and Univ Avignon, Eng/o)
* malaysia (Cheng Ming Guo, Universiti Sains Malaysia, Eng/a)
(details for Fr and It to follow)

Researchers/research groups hoping to return results for English within the next two months: Ted Pederson (California Polytech State Univ), Roberto Basili (Rome), Paul Rayson (Lancaster Univ), Mark Stevenson (Univ Sheffield).

English Pilot SENSEVAL: overview.

Adam Kilgarriff
ITRI
University of Brighton

August 28, 1998

[*] Auspices: ACL-SIGLEX, EURALEX, ELSNET, SPARKLE, ECRAN
Thanks: EPSRC, OUP, CUP, AWL, ELRA, EC DG XIII

The process

- Decide we're doing it
- Announce/encourage other language exercises

For English:

- choose task
  - relation to POS-tagging
  - all-words or lexical-sample
- choose dictionary (permissions)
- choose corpus (permissions)

If lexical sample:

- build sampling frame
- select sample
- define **tasks** eg *float-n*

Gold standard

- funding
- find good people; terms and conditions
- software, data formats
- detailed policy (eg *yell a promise*)
- first pass
- second pass

Participants

- advertise/encourage
- What sorts of systems are they?

Data

- DRY, TRAIN and EVAL
- Input format
- Output format

- POS anomalies
  . . . enough to stabilise a big float rig
  Keith Noble also float fished steak
  Ian Stanier won with three chub on float fished maggot
- WordNet/other mappings

Scoring

- Theory, coding
- Admin, hiccups
- Analysis of results

Inter-tagger Agreement

(English)

Adam Kilgarriff
ITRI, University of Brighton

Structure

- The Upper Bound problem
- (its solution)
- What is ITA?
- Numbers

Upper bound problem: Laments

If people can't agree, we don't even know what it **means** to say the computer got it right

- Jorgensen (1990) 68%
- Gale Church Yarowsky (ACL, 1993)
  Of course, it is a fairly major step to redefine the problem ... we simply don't know what else to do ...
- Ng and Lee (ACL, 1996) 57% (but)
- Bruce and Wiebe (EMNLP-3, 1998)
- Véronis (here)

Solution ...

...make it higher

- cf. Samuelsson and Voutilainen (1997) (POS-tagging)
- use experts
- use best possible quality dictionary
- typos are not interesting
- resolution phase **OK**
- dictionary improvement **OK**

Do we care what amateurs say?

- over 90% or it's fool's gold
- all except dict improvement
- replicability – to follow

What is ITA?

lurks round corners and scuttles away...

- Typos
- Simple errors
  at *giant-n* **n-prop** for **teams**
  at *promise-n* (verbal) **vow** for **vown**
- Not enough context
- In the middle or both:
  the rabbits were trapped, skinned and thrown in the pot
- Different interpretations of dictionary entry:
  syntax *vs.* semantics
  definition *vs.* examples

(see next talk)

TASK	N	PERFECT	FINE	COARSE
accident-n	267	0.94	8	2
amaze-v	70	0.95	1	1
band-p	302	0.98	29	25
behaviour-n	279	0.96	3	2
bet-n	275	0.87	15	9
bet-v	117	0.84	9	4
bother-v	209	0.90	8	6
brilliant-a	229	0.79	10	8
bury-v	201	0.82	14	6
calculate-v	218	0.90	5	3
consume-v	186	0.93	6	4
deaf-a	122	0.97	5	5
derive-v	217	0.87	6	4
disability-n	159	0.93	3	2
excess-n	186	0.88	8	3
float-n	74	0.93	12	8
float-v	228	0.78	16	11
generous-a	226	0.72	6	6
giant-a	97	0.96	5	2
giant-n	117	0.65	7	3
hurdle-p	322	0.90	11	8
invade-v	206	0.84	6	3
knee-n	250	0.97	22	12
modest-a	269	0.66	9	3
onion-n	213	0.92	4	4
promise-n	113	0.84	8	4
promise-v	224	0.84	6	3
rabbit-n	221	0.92	8	6
sack-n	82	0.98	11	9
sack-v	178	0.98	4	4
sanction-p	431	0.93	7	6
scrap-n	156	0.93	14	8
scrap-v	186	0.96	3	2
seize-v	259	0.89	11	9
shake-p	356	0.93	36	30
shirt-n	184	0.93	8	6
slight-a	218	0.99	6	3
steering-n	176	0.94	5	4
wooden-a	196	0.99	4	4
	(8438	0.89)		

AA-HEADER	diane	glennis	guy	john	lucy	ramesh	hector
accident-n	0.99	-	-	0.99	-	-	0.99
amaze-v	0.96	-	-	-	-	-	1.00
band-p	0.99	-	-	0.99	-	-	0.99
behaviour-n	1.00	-	-	-	-	1.00	0.97
bet-n	0.97	-	-	-	0.99	-	0.99
bet-v	1.00	-	-	-	0.97	-	0.95
bitter-p	-	0.98	-	-	-	0.98	0.95
bother-v	-	-	-	-	0.99	0.96	0.98
brilliant-a	-	0.95	-	-	-	0.98	0.95
bury-v	0.97	-	-	-	0.98	-	0.96
calculate-v	0.96	0.98	-	-	-	-	0.95
consume-v	0.99	-	-	-	0.99	-	0.96
deaf-a	-	0.99	-	0.98	-	-	0.99
derive-v	-	0.93	0.98	-	-	-	0.97
disability-n	-	0.99	1.00	-	-	-	0.96
excess-n	-	0.95	0.98	-	-	-	0.97
float-n	-	0.94	-	0.98	-	-	0.99
float-v	-	0.97	-	0.95	-	-	0.97
floating-a	-	-	-	-	-	-	0.98
generous-a	0.94	-	-	0.88	-	-	0.96
giant-a	1.00	-	-	1.00	-	-	1.00
giant-n	0.97	-	-	0.99	-	-	0.99
hurdle-p	0.99	-	-	0.97	-	-	0.98
invade-v	-	-	-	0.97	-	0.96	0.96
knee-n	-	-	-	0.97	-	0.98	0.99
modest-a	0.97	-	-	-	-	0.96	0.94
onion-n	0.99	-	-	-	-	0.97	0.95
promise-n	-	0.97	-	-	0.99	-	0.96
promise-v	-	0.98	-	-	0.96	-	0.96
rabbit-n	-	0.95	-	-	0.96	-	0.95
sack-n	1.00	-	-	-	0.98	-	1.00
sack-v	0.99	-	-	-	1.00	-	0.99
sanction-p	0.99	-	-	-	0.98	-	0.99
scrap-n	1.00	-	-	-	0.97	-	0.98
scrap-v	1.00	-	-	-	0.99	-	0.99
seize-v	-	0.96	0.98	-	-	-	0.95
shake-p	1.00	-	0.98	-	-	-	0.98
shirt-n	-	1.00	0.97	-	-	-	1.00
slight-a	-	-	1.00	0.99	-	-	1.00
steering-n	0.99	0.98	-	-	-	-	0.99
wooden-a	1.00	0.99	-	-	-	-	1.00
zz-eval	0.99	0.97	0.98	0.97	0.98	0.97	0.97

Two-way inter-tagger agreement

Coarse-grained, minimal scoring
averaged across x|y and y|x

HEADER	diane	glennis	guy	john	lucy	ramesh	hector
diane	1						0.94
glennis	0.96	1					0.92
guy	0.96	0.95	1				0.95
john	0.91	0.93	0.99	1			0.95
lucy	0.95	0.91	-	-	1		0.94
ramesh	0.93	0.91	-	0.95	0.93	1	0.93
GOLD	0.99	0.97	0.98	0.97	0.98	0.97	0.97

More than One Sense Per Discourse

Robert Krovetz
NEC Research Institute
Princeton, NJ 08540
krovetz@research.nj.nec.com

Abstract

Previous research has indicated that when a polysemous word appears two or more times in a discourse, it is extremely likely that they will all share the same sense [Gale et al. 92]. However, those results were based on a coarse-grained distinction between senses (e.g, *sentence* in the sense of a 'prison sentence' vs. a 'grammatical sentence'). We report on an analysis of multiple senses within two sense-tagged corpora, Semcor and DSO. These corpora used WordNet for their sense inventory. We found significantly more occurrences of multiple-senses per discourse than reported in [Gale et al. 92] (33% instead of 4%). We also found classes of ambiguous words in which as many as 45% of the senses in the class co-occur within a document. We discuss the implications of these results for the task of word-sense tagging and for the way in which senses should be represented.

1 Introduction

When a word appears more than once in a discourse, how often does it appear with a different meaning? This question is important for several reasons. First, the interaction between lexical semantics and discourse provides information about how word meanings relate to a larger context. In particular, the interaction provides a better understanding of the types of inferences involved. Second, by looking at word senses that systematically co-occur within a discourse we get a better understanding of the distinction between homonymy and polysemy (unrelated vs. related word senses).[1] Word senses that co-occur are more likely to be related

[1] For example, *race* is homonymous in the sense of 'human race' vs. 'horse race'. *Door* is polysemous in the contexts 'paint the door' vs. 'go through the door'.

than those that are not. Finally, the question is important for word sense tagging. If a word appears with only one meaning in a discourse then we can disambiguate only one occurrence and tag the rest of the instances with that sense.

Prior work on the number of senses per discourse was reported in [Gale et al. 92]. Their work was motivated by their experiments with word sense disambiguation. They noticed a strong relationship between discourse and meaning and they proposed the following hypothesis: *When a word occurs more than once in a discourse, the occurrences of that word will share the same meaning.*

To test this hypothesis they conducted an experiment with five subjects. Each subject was given a set of definitions for 9 ambiguous words and a total of 82 pairs of concordance lines for those words. The subjects were asked to determine for each pair whether they corresponded to the same sense or not. The researchers selected 54 pairs from the same discourse and 28 were used as a control to force the judges to say they were different. The control pairs were selected from different discourses and were checked by hand to assure that they did not use the same sense. The result was that 51 of the 54 pairs were judged to be the same sense (by a majority opinion). Of the 28 control pairs, 27 were judged to be different senses. This gave a probability of 94% (51/54) that two ambiguous words drawn from the same discourse will have the same sense. [Gale et al. 92] then assumed that there is a 60/40 split between unambiguous/ambiguous words, so there is a 98% probability that two word occurrences in the same discourse will have the same sense.

[Gale et al. 92] suggested that these results could be used to provide an added constraint for improving the performance of word-sense disambiguation algorithms. They also proposed that it be used to help evaluate sense tagging. Only one instance of the word in a discourse would need to be tagged and the remaining instances could be tagged automatically with the same sense. This would provide a much larger set of training instances, which is a central problem for disambiguation.

In our own experiments with disambiguation we found a number of instances where words appeared in the same document with *more* than one meaning [Krovetz and Croft 92]. These observations were based on experiments with two corpora used in information retrieval. One corpus consisted of titles and abstracts from Communications of the ACM (a Computer Science journal). The other corpus consisted of short articles from TIME magazine. In the CACM corpus a word rarely appeared more than once in a document (since the documents were so short). However, in the TIME corpus we found a number of cases where words appeared in the same document with more than one meaning. A sample of these words is given below:

party dinner party / political party

headed headed upriver / headed by

great great grandson / Great Britain
great Irishmen / Great Britain

park Industrial park / Dublin's park
Industrial park / parking meter

line a line drawn by the U.S. / hot line

We even found one instance in which five different senses of a word occurred within the same document: 'mile long cliff *face*', 'difficulties ... is *facing* because', 'in the *face* of temptations', 'about *face*', and 'his pavilion *facing* lovely west lake'[2]

[Gale et al. 92]'s hypothesis raises the question: What is a *sense*? Most of the work on sense-disambiguation has focused on meanings that are unrelated, the so-called 'Bank model' (river bank vs. savings bank). But in practice word senses are often related. Unrelated senses of a word are *homonymous* and related senses are termed *polysemous*.[3] In [Gale et al. 92]'s experiments they asked the subjects to determine whether the pairs of concordance lines exhibited the same sense or not. But human judgement will vary depending on whether the senses are homonymous or polysemous [Panman 82]. People will often agree about the sense of a word in context when the senses are unrelated (e.g., we expect that people will reliably tag 'race' in the sense of a horse race vs. human race), but people will disagree when the senses are related.

The disagreement between individuals about polysemous senses might be considered an impediment, but we prefer to view it as a source of data. We can use the judgements to help distinguish homonymous from polysemous senses. When the judgements are *systematically* inconsistent, we predict that the senses will be polysemous. In other words, the inconsistency in human judgement (with respect to determining the meaning of a word in context) can be viewed as a feature rather than a bug.

In addition, there are a variety of tests to help establish word sense identity. For example, we can conjoin two senses and note the anomaly (zeugma): "The newspaper fired its employees and fell off the table" [Cruse 86]. We can also determine whether a word is a member of a class that is systematically ambiguous (e.g., language/people or object/color - see [Krovetz 93]).

[2] These examples illustrate a difference from other work on word meanings. Most of that work has not considered any morphological variants for a word or differences across part of speech.

[3] The word *polysemy* is itself polysemous. In general usage it is a synonym for lexical ambiguity, but in linguistics it refers to senses that are related.

[Gale et al. 92]'s hypothesis also raises the question: What is a *discourse*? Is it a paragraph, a newspaper article, a document that is about a given topic, or something else? How do the concepts of discourse and topic relate to each other? Research on topic segmentation [Hearst 97] and work on text coherence [Morris and Hirst 91] addresses this question. We can't provide an answer to how this work affects [Gale et al. 92]'s hypothesis, but the question of what constitutes a discourse is central to its testability.

This paper is concerned with the first question we raised - how does word sense identity affect [Gale et al. 92]'s results? In particular, what happens if we consider the distinction between homonymy and polysemy? We conducted experiments to determine whether [Gale et al. 92]'s hypothesis would hold when applied to finer grained sense distinctions. These experiments are described in the following section.

2 Experiments

Our experiments used two sense-tagged corpora, *Semcor* [Miller et al. 94] and *DSO* [Ng and Lee 96]. Both of these corpora used WordNet as a basis for the sense inventory [Miller 1990]. WordNet contains a large number of words and senses, and is comparable to a good collegiate dictionary in its coverage and sense distinctions. Semcor is a semantic concordance in which all of the open class words[4] for a subset of the Brown corpus[5] were tagged with the sense in WordNet. The DSO corpus is organized differently from Semcor. Rather than tag all open-class words, it consists of a tagging of 191 highly ambiguous words in English within a number of files. These files are drawn from the Brown corpus and the Wall Street Journal. The 191 words are made up of 121 nouns and 70 verbs.

We conducted experiments to determine how often words have more than one meaning per discourse in the two sense-tagged corpora. This was defined as more than one WordNet sense tag in a file from the Brown corpus (for Semcor) and in a file from either the Brown Corpus or the Wall Street Journal for DSO.

For Semcor we wrote a program to identify all instances in which a tagged word occurred in a file from the Brown corpus with more than one sense. The program determined the potential ambiguity of these words (the number of senses they had in WordNet) as well as the actual ambiguity (the number of senses for those words in Semcor). We then computed the proportion of the ambiguous words within the corpus that had more than one sense in a document.

For the DSO corpus we determined how many of the tagged words had more than one

[4] Nouns, verbs, adjectives, and adverbs.

[5] The Brown corpus consists of 500 discourse fragments of 2000 words, each.

sense in a document. We also determined how many documents contained an instance of the tagged word with more than one sense.

3 Results

The statistics for the experiment are given in Table 1. We indicate the number of unique words with a breakdown according to part of speech. We also show the number of words that have more than one sense in WordNet (potential ambiguity) and the number that have more than one sense in the corpus (actual ambiguity). Finally, we indicate the number of words that have more than one sense per discourse.

The statistics provide a strong contrast with the results from [Gale et al. 92]. About 33% of the ambiguous words in the corpus had multiple senses within a discourse. There was no difference in this respect for the different parts of speech.

However, the statistics do show significant differences between the different parts of speech with regard to potential vs. actual ambiguity. The proportion of ambiguous words in WordNet [potential ambiguity] was 47% for nouns, 66% for verbs, and 63% for adjectives. The proportion of potentially ambiguous words that were found to be ambiguous in the corpus was 41%, 50% and 18% for nouns, verbs, and adjectives (respectively). We do not have any explanation for why the actual ambiguity for adjectives is so low.

We also examined words that were ambiguous with regard to part-of-speech. There were 752 words in Semcor that were ambiguous between noun and verb. Of these words, 267 (36%) appeared in a document in both forms. There were 182 words that were ambiguous between noun and adjective. Of these words, 82 (45%) appeared in a document in both forms.

The results with the DSO corpus support the findings with Semcor. *All* of the 191 words were found to occur in a discourse with more than one sense. On average, 39% of the files containing the tagged word had occurrences of thè word with different senses.

4 Analysis

When two senses co-occur in a discourse it is possible that the co-occurrence is accidental. We therefore examined those senses that co-occured in four or more files (for nouns) and three or more files (for verbs and adjectives).

For nouns, the systematic sense co-occurrences were primarily due to logical polysemy [Apresjan 75], [Pustejovsky 95] or to general/specific sense distinctions. A sample of these

	Nouns	Verbs	Adj
Word Types	8451	3296	1521
Potential ambiguity	4016	2161	962
Actual ambiguity	1659	1089	169
Multiple Sense/Discourse	517	365	55

Table 1: Statistics on multiple-senses within a discourse for Semcor. *Potential ambiguity* refers to the number of unique words that have more than one sense in WordNet. *Actual ambiguity* is the number of those words that were found to have more than one sense within the tagged corpus.

co-occurrences is given below[6]:

Logical Polysemy

agent/entity (city, school, church)
meal/event (dinner)
language/people (Assyrian, English)
figure/ground (door)
result/process (measurement)
metonymy (sun, diameter)

General/Specific

day (solar *day*/mother's *day*)
question (the *question* at hand/ask a *question*)
man (race of *man*/bearded *man*)

The figure/ground ambiguity refers to *door* as a physical object or to the space occupied by the door. The metonymic ambiguity for *sun* refers to the physical object as opposed to the rays of the sun. For *diameter* we can refer to the line or to the length of the line.

For verbs, the sense co-occurrences were more difficult to characterize. They generally seemed like active/passive distinctions. For example:

see 'We saw a number of problems' (recognize)
'We saw the boat' (perceive)

[6] Some of the examples occurred in less than four files, but we mention them because they help to illustrate the members of the class.

know 'know a fact' (be-convinced-of)
'know the time' (be-aware-of)

remember 'remember to bring the books' (keep-in-mind)
'remember when we bought the books' (recollect)

For adjectives the different senses reflect either differing dimensions, or absolute/relative distinctions:

old not young vs. not new

long spatial vs. temporal

little not big vs. not much

same identical vs. similar

The noun/verb ambiguities often reflected a process/result difference (e.g., *smile*, *laugh*, or *name*). The noun/adjective ambiguities represent a number of systematic classes:

nationality or religion British, German, Catholic, American, Martian (!)

belief humanist, liberal, positivist

made-of chemical, liquid, metal

gradable-scale quiet, young, cold

We note that there are some cases where multiple senses *might* have been identified, but WordNet was not consistent in the distinctions in meaning. For example, *dinner* has the meal vs. event distinction, but the same ambiguity was not represented for *lunch* or *breakfast*. *Assyrian*, and *English* have the language/people distinction, but these senses were not provided for *Dutch* or *Korean*. These omissions are not a criticism against WordNet per se - dictionaries are not designed to contain systematic sense distinctions whenever we have logical polysemy. In our work with the Longman Dictionary [Procter 78] we noticed a number of cases where sense distinctions were not made systematically. These inconsistencies are a reflection of human judgement with regard to polysemy.

The polysemous relations we found for isolated words were also found for lexical phrases. Although phrases usually have only one meaning,[7] we found instances in which they occurred with more than one sense within a discourse. Out of eight ambiguous lexical phrases in Semcor,[8] three occurred with more than one sense in a discourse. These phrases were: *United States* (country vs. government), *interior design* (branch of architecture vs. occupation), and *New York* (city vs. state). The first two instances are similar to other classes of logical polysemy that have been reported in the literature. The country vs. government distinction is akin to the difference between *white house* as a physical entity vs. as an agent ('He entered the White House' vs. 'The White House dismissed the chief prosecutor'). The ambiguity between fields of knowledge and occupations is also common. Although lexical phrases have less ambiguity than isolated words, we observe that the different senses can still co-occur.

The co-occurrence of multiple senses within a discourse can be used as evidence for lexical semantic relations, and to help distinguish homonymy from polysemy. So *quack* as a noun and as a verb are related in the sense of a sound made by a duck, but not in the sense of a bad doctor. This is akin to gravity/gravitation being related in the sense of 'the force of gravity', but not with regard to the 'gravity of the offense'. In our earlier work we established links between senses in the dictionary by looking for words which occurred in their own definition, but with a different part of speech. We in essence treated dictionary definitions as a small "discourse" (we can even find deictic relationships between dictionary definitions - see [Krovetz 93]). The hypothesis is that if senses co-occur within a discourse they will be related even if they differ in part-of-speech. For example, we would predict that *paint* as a noun and as a verb will co-occur in a discourse much more often than *train* as a noun and as a verb.

We can learn about lexical semantic relations by examining dictionary definitions of related senses. For example, the relationship between *dust* as a noun and as a verb can be one of covering or removing. The dictionary tells us that it has both meanings.

The biggest problem we encountered in our analysis was the number of tagged files. We wanted to ensure that the sense co-occurrences were not simply an accident, so we looked for sense pairs that co-occurred in several files. But the existing tagged corpora are not large enough to get reliable statistics. *Dust* as a verb only appears twice out of the 106,000 tagged word forms in Semcor. This is not often enough to get statistics about co-occurrence with a noun, much less co-occurrence with specific senses.

[7] This generalization is not true for phrasal verbs (verb-particle constructions).

[8] These phrases are all nouns. We also noticed senses of verbs that co-occurred. However, it is especially difficult to analyze phrasal lexemes because they occur less frequently than isolated words. Co-occurrences for particular senses are even more infrequent.

5 Conclusions and Future Work

[Gale et al. 92]'s hypothesis is probably correct for homonymous senses. It is unlikely that a document which mentions *bank* in the sense of a river bank will also use it in the sense of a savings bank. However, even with homonymous senses, we expect there will be certain cases that will predictably co-occur. For example, in legal documents *support* in the sense of *child support* can co-occur with *support* in the sense of supporting an argument. The work reported in this paper shows that the hypothesis is not true with regard to senses that are polysemous.

We do not want to give the impression that the distinction between homonymy and polysemy is straightforward. It is not. In practice the differences in meaning are not always clear. But that does not mean that the distinction between homonymy and polysemy is vacuous. We gain a better understanding of the difference by looking at systematic classes of ambiguity. Another set of semantically tagged files was just released.[9] These files will allow us to examine a larger number of words in which the multiple senses co-occurrences are systematic.

Our results indicate that we cannot simply adopt [Gale et al. 92]'s suggestion that we disambiguate one occurrence of a word in a discourse and then assign that sense to the other occurrences. However, we *can* leverage the systematic classes of ambiguity. If a word appears in a discourse and there are senses of that word that are systematically polysemous, we can attempt to tag the other occurrences in the discourse in light of this ambiguity. In the future we will examine rules associated with classes of polysemous words that will allow these occurrences to be tagged.

Acknowledgements

I am grateful to David Lebeaux and Christiane Fellbaum for their comments on this paper.

References

[Apresjan 75] Apresjan Ju, "Regular Polysemy", *Linguistics*, Vol. 142, pp. 5-32, 1975.

[Cruse 86] Cruse David, *Lexical Smantics*, Cambridge University Press, 1986.

[9] Brown2, which consists of an additional 83 files.

[Gale et al. 92] Gale William, Kenneth Church, and David Yarowsky, "One Sense Per Discourse", in *Proceedings of the ARPA Workshop on Speech and Natural Language Processing*, pp. 233–237, 1992.

[Hearst 97] Hearst Marti, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages", *Computational Linguistics*, Vol. 23(1), pp. 33–64, 1997.

[Krovetz and Croft 92] Krovetz Robert and W. Bruce Croft, "Lexical Ambiguity and Information Retrieval", *ACM Transactions on Information Systems*, pp. 145–161, 1992.

[Krovetz 93] Krovetz Robert, "Sense Linking in a Machine-Readable Dictionary", in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 330–332, 1993.

[Miller 1990] Miller George, "WordNet: An on-line Lexical Database", *International Journal of Lexicography*, Vol. 3(4), pp. 235-312, 1990.

[Miller et al. 94] Miller George, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert Thomas, "Using a Semantic Concordance for Sense Identification", in *Proceedings of the ARPA Human Language Technology Workshop*, 1994.

[Morris and Hirst 91] Morris Jane and Graeme Hirst, "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", *Computational Linguistics*, Vol. 17(1), pp. 21–48, 1991.

[Ng and Lee 96] Ng Hwee Tou and Hian Beng Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach", in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 40–47, 1996

[Panman 82] Panman Otto, "Homonymy and Polysemy", *Lingua*, Vol. 58, pp. 105–136, 1982

[Procter 78] Procter Paul, *Longman Dictionary of Contemporary English*, Longman, 1978.

[Pustejovsky 95] Pustejovsky James, *The Generative Lexicon*, MIT Press, 1995

[Yarowsky 92] Yarowsky David, "Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in *Proceedings of the 14th Conference on Computational Linguistics, COLING-92*, pp. 454–450, 1992.

SENSEVAL: The CL Research Experience

Ken Litkowski

Abstract: CL Research achieved a reasonable level of performance in the final SENSEVAL word-sense disambiguation evaluation, with a overall fine-grained score of 52 percent for recall and 56 percent for precision on 93 percent of the 8,448 texts. These results were significantly affected by time constraints; results from the training data and initial perusal of the submitted answers strongly suggest an additional 15 percent for recall, 10 percent for precision, and coverage of nearly 100 percent could have been achieved without looking at the answer keys. These results were achieved with an almost complete reliance on syntactic behavior, as time constraints severely limited the opportunity for incorporation of various semantic disambiguation strategies. The results were achieved primarily through the performance of (1) a robust and fast ATN-style parser producing parse trees with annotations on nodes, (2) the use of the DIMAP dictionary creation and maintenance software (via conversion of the HECTOR dictionary files), and (3) the strategy for analyzing the parse trees with the dictionary data. Several potential avenues for increasing performance were investigated briefly during development of the system and suggest the likelihood of further improvements. SENSEVAL has provided an excellent testbed for the development of practical strategies for analyzing text. These strategies are now being expanded to include (1) parsing of dictionary definitions in MRDs to create entries like those used in SENSEVAL (and simultaneously, creating semantic network links), (2) analysis of corpora to extract dictionary information to create entries, and (3) extraction of information for creation of knowledge bases.

Combining heterogeneous knowledge (upc-ehu)

Eneko Agirre	Basque Country University
Jordi Atserias	
Lluis Padró	Polytechnic University of Catalonia
German Rigau	

Knowledge needed (McRoy 1992; Hirst 1987)

- part of speech
- morphology
- syntactic clues and collocational information
- selectional restrictions
- relationship with other words in context
- knowledge of context (topic and domain)
- general inference

Potential knowledge sources

- Syntactic
  - pos taggers
  - multiword term recognizers
  - ...
- Semantic
  - ontologies
  - dictionaries
  - corpora

Combination

- Identify independent knowledge sources
- Combination of classifiers using unweighted voting
  Machine Learning (Dietterich, 1997)

Upc-ehu systems

- Preliminary implementation

  pos tagger (Padró, 1998)

  multiword recognizer

  ontologies ⇒ hierarchy of WordNet

  dictionaries ⇒ definitions in WordNet

  corpora ⇒ cooccurrences on Hector

Upc-ehu systems

- Ontologies:

  Conceptual Density on WordNet

  (Agirre & Rigau, 1996; Agirre, forthcoming)

- uses hierarchical knowledge

Upc-ehu systems

- Dictionaries:

  definitions + synsets from WordNet

  (Rigau & Agirre, 1997)

  - sense ordering
  - word match
  - topic match

Upc-ehu systems

- Corpora:

  Decision lists on mutual information for cooccurrences in Hector corpus

  (Yarowsky, 1994)

Upc-ehu systems

- Focused on nouns only
- UNSUPERVISED: Upc-ehu-uns
  combine 4 heuristics with unweighted votes
  yields the result on WordNet synsets
- SUPERVISED: Upc-ehu-sup:
  translate WordNet synsets to Hector
  combine, giving winning weight to dlists.

Conclusions

- Preliminary system (more ambitious to come)
- Pos tagger and multiword recognizer (?)
- Little human resources
- No time to test or fit (WN-Hector mapping)
  ⇒ test run only
- No sense in tagging WordNet and translating to Hector (upper bound?)

A = no training data

shake 700001:

What at the end of forty years, eh?"

Here he <tag "shake-v_shake/1/1_move"> shook the bag again.

```
A clres.fix                          shake-v_shake/1/1_move
S durham.fix                         shake-v_shake/1/1_move
S ets-pu.fix                         shake-v_shake/1/1_move
S grling-sdm.fix                     shake-v_shake/1/1_move
O hopkins.fix                        shake-v_shake/1/2_tremble
S korea.fix                          shake-v_shake/1/3_head
A malaysia.fix                       shake-n_shake/1/9_movement
A manitoba.dl.dictonly.fix           shake-v_shake/1/1_move / 1
A manitoba.dl.fix                    shake-v_shake/1/2_tremble / 0.695364
S manitoba.ks.fix                    shake-v_shake/1/4_hand
A suss.fix                           shake-v_shake/1/1_move
S tilburg.fix                        shake-v_shake/1/1_move
A xeroxceli.fix                      shake-v_shake/1/1_move
S commonest                          shake-v_shake/1/3_head
```

shake 700002:

They'll just make you over in the studio."

Martha <tag "shake-v_shake/1/3_head"> shook her head and tossed the letter on to the table.

```
A clres.fix                          shake-v_shake/1/3_head
S durham.fix                         shake-v_shake/1/3_head
S ets-pu.fix                         shake-v_shake/1/3_head
S grling-sdm.fix                     shake-v_shake/1/3_head
O hopkins.fix                        shake-v_shake/1/3_head
S korea.fix                          shake-v_shake/1/3_head
A malaysia.fix                       shake-v_shake/1/3_head
A manitoba.dl.dictonly.fix           shake-v_shake/1/3_head / 0.470178
A manitoba.dl.fix                    shake-v_shake/1/3_head / 0.470178
S manitoba.ks.fix                    shake-v_shake/1/3_head
A suss.fix                           shake-v_shake/1/3_head
S tilburg.fix                        shake-v_shake/1/3_head
A xeroxceli.fix                      shake-v_shake/1/1.1_clean
S commonest                          shake-v_shake/1/3_head
```

shake 700003:

The majority of opinion reports from the SD and other agencies of the regime reaching the Nazi leadership point nevertheless towards conclusions about the impact on morale similar to those we have witnessed for the Schweinfurt area.

And Goebbels's own diary jottings leave little doubt that he thought morale was severely <tag "shake-v_shake/1/7_ideas"> shaken by the bombing, and the will to resist potentially weakened.

```
A clres.fix                          shake-v_shake/1/7_ideas
S durham.fix                         shake-v_shake/1/6_disturb
S ets-pu.fix                         shake-v_shake/1/6_disturb
S grling-sdm.fix                     shake-a_shaken_troubled
O hopkins.fix                        shake-v_shake/1/6_disturb
S korea.fix                          shake-a_shaken_troubled
A malaysia.fix                       shake-a_shaken_troubled
A manitoba.dl.dictonly.fix           shake-v_shake/1/6_disturb / 0.332139
A manitoba.dl.fix                    shake-a_shaken_troubled / 0.565562
S manitoba.ks.fix                    shake-a_shaken_troubled
A suss.fix                           shake-a_shaken_troubled
```

```
S tilburg.fix                        shake-a_shaken_troubled
A xeroxceli.fix                      shake-v_shake/1/7_ideas
S commonest                          shake-v_shake/1/3_head
```

shake 700004:

Morning newspapers are regularly sold out by eight o'clock.

Old puppet institutions have been disbanded or <tag "shake-v_shake_up/5/2_up"> shaken up.

```
A clres.fix                          shake-v_shake_up/5/2_up
S durham.fix                         shake-v_shake_up/5/2_up
S ets-pu.fix                         shake-v_shake_up/5/3_emotion
S grling-sdm.fix                     shake-v_shake_up/5/2_up
O hopkins.fix                        shake-v_shake_up/5/2_up
S korea.fix                          shake-v_shake_up/5/2_up
A malaysia.fix                       shake-a_shaken_troubled
A manitoba.dl.dictonly.fix           shake-v_shake/1/3_head / 0.310882
A manitoba.dl.fix                    shake-v_shake/1/2_tremble / 0.347378
S manitoba.ks.fix                    shake-a_shaken_troubled
A suss.fix                           shake-a_shaken_troubled
S tilburg.fix                        shake-v_shake_up/5/2_up
S commonest                          shake-v_shake/1/3_head
```

shake 700005:

`Looks like you had a letter for him ."

Rain <tag "shake-v_shake/1/3_head"> shook her head.

```
A clres.fix                          shake-v_shake/1/3_head
S durham.fix                         shake-v_shake/1/3_head
S ets-pu.fix                         shake-v_shake/1/3_head
S grling-scm.fix                     shake-v_shake/1/3_head
O hopkins.fix                        shake-v_shake/1/3_head
S korea.fix                          shake-v_shake/1/3_head
A malaysia.fix                       shake-v_shake/1/3_head
A manitoba.dl.dictonly.fix           shake-v_shake/1/3_head / 0.811323
A manitoba.dl.fix                    shake-v_shake/1/3_head / 1
S manitoba.ks.fix                    shake-v_shake/1/3_head
A suss.fix                           shake-v_shake/1/3_head
S tilburg.fix                        shake-v_shake/1/3_head
A xeroxceli.fix                      shake-v_shake/1/2_tremble
S commonest                          shake-v_shake/1/3_head
```

shake 700006:

Mr Krenz, a former head of the Communist Youth Movement and long the heir to the former leader, Erich Honecker, had conspired to topple Mr Honecker after the mass exodus of East Germans and huge demonstrations at home made it clear things had to change.

He opened the Berlin Wall and the border to let his people travel; he promised free, multi-party elections and eventually agreed to abolish the Communists' constitutional right to political control.

But, for all his efforts, he never gained credibility, and was unable to <tag "shake-v_shake_off/3/1_off"> shake off charges that he rigged the last elections, or take back his public support for the massacre in Tiananmen Square.

```
A clres.fix                    shake-v_shake_off/3/1_off
S durham.fix                   shake-v_shake_off/3/1_off
```

```
S ets-pu.fix                    shake-v_shake_off/3/1_off
S grling-sdm.fix                shake-v_shake_off/3/1_off
O hopkins.fix                   shake-v_shake_off/3/1_off
S korea.fix                     shake-v_shake_off/3/1_off
A malaysia.fix                  shake-v_shake/1/7_ideas
A manitoba.dl.dictonly.fix      shake-v_shake_up/5/2_up / 0.384397
A manitoba.dl.fix               shake-v_shake_off/3/1_off / 0.483294
S manitoba.ks.fix               shake-v_shake_up/5/2_up
A suss.fix                      shake-v_shake_off/3/1_off
S tilburg.fix                   shake-v_shake_off/3/1_off
A xeroxceli.fix                 shake-v_shake/1/1_move
S commonest                     shake-v_shake/1/3_head
```

shake 700007:

I managed to get down the last two words of the preceding paragraph before my stomach over-boiled into my mouth.

I rushed down the dark passage to the lavatory with both hands at my face.

I do not ever recall being quite as sick and <tag "shake-a_shaken_troubled"> shaken as I was then, about an hour and a half ago.

```
A clres.fix                  shake-a_shaken_troubled
S durham.fix                 shake-v_shake/1/6_disturb
S ets-pu.fix                 shake-v_shake/1/1_move
S grling-sdm.fix             shake-v_shake/1/6_disturb
O hopkins.fix                shake-v_shake/1/2_tremble
S korea.fix                  shake-a_shaken_troubled
A malaysia.fix               shake-a_shaken_troubled
A manitoba.dl.dictonly.fix   shake-v_shake/1/3_head / 0.747414
A manitoba.dl.fix            shake-v_shake/1/2_tremble / 1
S manitoba.ks.fix            shake-a_shaken_troubled
A suss.fix                   shake-a_shaken_troubled
S tilburg.fix                shake-v_shake/1/6_disturb
S commonest                  shake-v_shake/1/3_head
```

shake 700008:

For the second time the rebels have got into the wealthy areas and the army hasn't been able to push them out until they were ready to leave."

The guerrillas' first urban offensive, which has lasted three weeks so far and shows no sign of ending, has <tag "shake-v_shake/1/6_disturb"> shaken a city lulled by the official propaganda.

```
A clres.fix                    shake-a_shaken_troubled
S durham.fix                   shake-v_shake/1/6_disturb
S ets-pu.fix                   shake-v_shake/1/1_move
S grling-sdm.fix               shake-v_shake/1/6_disturb
O hopkins.fix                  shake-v_shake/1/1_move
S korea.fix                    shake-v_shake/1/6_disturb
A malaysia.fix                 shake-a_shaken_troubled
A manitoba.dl.dictonly.fix     shake-v_shake/1/3_head / 0.599458
A manitoba.dl.fix              shake-v_shake/1/2_tremble / 0.751752
S manitoba.ks.fix              shake-a_shaken_troubled
A suss.fix                     shake-a_shaken_troubled
S tilburg.fix                  shake-v_shake/1/6_disturb
A xeroxceli.fix                shake-v_shake/1/7_ideas
S commonest                    shake-v_shake/1/3_head
```

shake 700009:

From the recesses of her memory emerged the stories she had half-heard and loyally ignored all her life, of subnormal or afflicted members of the royal lineage who had lived their sad lives in obscurity.
Wood Farm, she recalled, had been a home for one of them; the place she had felt hallowed by her own happiness was now part of the sinister pattern.

She <tag "shake-v_shake/1/3_head"> shook her head violently to shut out the notion, and grasped the door-knob for support as she swayed off-balance.

```
A clres.fix                    shake-v_shake/1/3_head
S durham.fix                   shake-v_shake/1/3_head
S ets-pu.fix                   shake-v_shake/1/3_head
S grling-sdm.fix               shake-v_shake/1/3_head
O hopkins.fix                  shake-v_shake/1/3_head
S korea.fix                    shake-v_shake/1/3_head
A malaysia.fix                 shake-v_shake/1/3_head
A manitoba.dl.dictonly.fix     shake-v_shake/1/3_head / 1
A manitoba.dl.fix              shake-v_shake/1/3_head / 1
S manitoba.ks.fix              shake-v_shake/1/3_head
A suss.fix                     shake-v_shake/1/3_head
S tilburg.fix                  shake-v_shake/1/3_head
A xeroxceli.fix                shake-v_shake/1/1_move
S commonest                    shake-v_shake/1/3_head
```

onion 700001:

They had obviously simply persuaded others to go through this part of their therapy for them.

`I want salt and vinegar, chilli beef and cheese and <tag "onion-n_onion//1_veg"> onion!" said Maisie.

```
O avignon.fix                  onion-n_onion//1_veg / 0.65
                               onion-n_onion//1_veg / 0.080
                               **ANY**-*ANY*_UNASSIGNABLE_U / 0.080
A clres.fix                    onion-n_onion//1_veg
S durham.fix                   onion-n_onion//1_veg
S ets-pu.fix                   onion-n_onion//1_veg
S grling-sdm.fix               onion-n_onion//1_veg
O hopkins.fix                  onion-n_onion//1_veg
S korea.fix                    onion-n_onion//1_veg
A malaysia.fix                 onion-n_onion//1_veg
A manitoba.dl.dictonly.fix     onion-n_onion//1_veg / 0.404656
A manitoba.dl.fix              onion-n_onion//1_veg / 0.451312
S manitoba.ks.fix              onion-n_onion//1_veg
A suss.fix                     onion-n_onion//1_veg
S tilburg.fix                  onion-n_onion//1_veg
A upc-ehu-su.fix               onion-n_onion//1_veg / 9
                               onion-n_onion//2_plant / 1.9
A upc-ehu-un.fix               onion-n_onion//1_veg / 4
                               onion-n_onion//2_plant / 1.9
S commonest                    onion-n_onion//1_veg
```

onion 700002:

`Or perhaps you'd enjoy a bratwurst omelette?"

Pale, Chay told the waiter to have the kalbsbratwursts parboiled for four minutes at simmer then to grill them and serve them with smothered fried <tag "onion-n_onion//1_veg"> onions and some Dijon mustard.

```
O avignon.fix                 onion-n_onion//1_veg / 0.67
                              onion-n_spring_onion_spring / 0.15
```

```
A clres.fix                  onion-n_onion//1_veg
S durham.fix                 onion-n_onion//1_veg
S ets-pu.fix                 onion-n_onion//1_veg
S grling-sdm.fix             onion-n_onion//1_veg
O hopkins.fix                onion-n_onion//1_veg
S korea.fix                  onion-n_onion//1_veg
A manitoba.dl.dictonly.fix onion-n_onion//1_veg / 0.251977
A manitoba.dl.fix            onion-n_onion//1_veg / 0.300843
S manitoba.ks.fix            onion-n_onion//1_veg
A ottawa.ret.fix             onion-n_onion//2_plant
A suss.fix                   onion-n_onion//1_veg
S tilburg.fix                onion-n_onion//1_veg
A upc-ehu-su.fix             onion-n_onion//1_veg / 3.66667
                             onion-n_onion//2_plant / 2.4
A upc-ehu-un.fix             onion-n_onion//1_veg / 3.66667
                             onion-n_onion//2_plant / 2.4
A xeroxceli.fix              onion-n_onion//2_plant
S commonest                  onion-n_onion//1_veg
```

onion 700003:

With the motor running, slowly add the oil until the mixture is the consistency of a thick mayonnaise.

Stir in the <tag "onion-n_onion//1_veg"> onion, add the salt and pepper or a little more lemon juice if required.

```
O avignon.fix                  onion-n_onion//1_veg / 0.97
A clres.fix                    onion-n_onion//1_veg
S durham.fix                   onion-n_onion//1_veg
S ets-pu.fix                   onion-n_onion//1_veg
S grling-sdm.fix               onion-n_onion//1_veg
O hopkins.fix                  onion-n_onion//1_veg
S korea.fix                    onion-n_onion//1_veg
A malaysia.fix                 onion-n_onion//1_veg
A manitoba.dl.dictonly.fix     onion-n_onion//1_veg / 0.440054
A manitoba.dl.fix              onion-n_onion//1_veg / 0.563236
S manitoba.ks.fix              onion-n_onion//1_veg
A suss.fix                     onion-n_onion//1_veg
S tilburg.fix                  onion-n_onion//1_veg
A upc-ehu-su.fix               onion-n_onion//1_veg / 9
                               onion-n_onion//2_plant / 1.73333
A upc-ehu-un.fix               onion-n_onion//1_veg / 4
                               onion-n_onion//2_plant / 1.73333
A xeroxceli.fix                onion-n_onion//2_plant
S commonest                    onion-n_onion//1_veg
```

onion 700004:

The huge browned turkey was placed in the centre of the table.

The golden stuffing was spooned from its breast, white dry breadcrumbs spiced with <tag "onion-n_onion//1_veg"> onion and parsley and pepper.

```
O avignon.fix                  onion-n_onion//1_veg / 0.66
                               onion-n_onion//1_veg / 0.080
                               **ANY**-*ANY*_UNASSIGNABLE_U / 0.080
A clres.fix                    onion-n_onion//1_veg
S durham.fix                   onion-n_onion//1_veg
S ets-pu.fix                   onion-n_onion//1_veg
S grling-sdm.fix               onion-n_onion//1_veg
O hopkins.fix                  onion-n_onion//1_veg
S korea.fix                    onion-n_onion//1_veg
A malaysia.fix                 onion-n_onion//1_veg
A manitoba.dl.dictonly.fix     onion-n_onion//1_veg / 0.292592
```

```
A manitoba.dl.fix              onion-n_onion//1_veg / 0.521744
S manitoba.ks.fix              onion-n_onion//1_veg
A suss.fix                     onion-n_onion//1_veg
S tilburg.fix                  onion-n_onion//1_veg
A upc-ehu-su.fix               onion-n_onion//1_veg / 4
                               onion-n_onion//2_plant / 2.02121
A upc-ehu-un.fix               onion-n_onion//1_veg / 4
                               onion-n_onion//2_plant / 2.02121
A xeroxceli.fix                onion-n_onion//2_plant
S commonest                    onion-n_onion//1_veg
```

onion 700005:

Ingredients:

12oz / 375g nince 1oz / 30ml vegetable or olive oil 2 medium <tag "onion-n_onion//1_veg"> onions, diced 1 green pepper, diced 3 stalks celery, sliced 1 tin (14oz / 400g) plum tomatoes 1tsp sugar Cayenne pepper to taste (at least 1 / 2 tsp) Salt, pepper Half a 14oz / 400g tin of red kidney beans, drained, or 7oz / 200g tin of sweetcorn, drained 1 jalapeno pepper, sliced (optional) For the cornbread: 4oz / 125g cornmeal (yellow coarse grind &dash. the Encona brand is widely available) 1oz / 30g plain flour 1 / 2 tsp salt 1tsp baking powder 1 egg 5oz / 150ml milk 1tbs vegetable oil 2oz / 60g grated cheese Method: In a saute pan, brown meat in oil; stir in onions, green pepper and celery.

```
O avignon.fix                  onion-n_onion//1_veg / 0.74
A clres.fix                    onion-n_onion//1_veg
S durham.fix                   onion-n_onion//1_veg
S ets-pu.fix                   onion-n_onion//1_veg
S grling-sdm.fix               onion-n_onion//1_veg
O hopkins.fix                  onion-n_onion//1_veg
S korea.fix                    onion-n_onion//1_veg
A manitoba.dl.dictonly.fix     onion-n_onion//1_veg / 0.237701
A manitoba.dl.fix              onion-n_onion//1_veg / 0.23263
S manitoba.ks.fix              onion-n_onion//1_veg
A suss.fix                     onion-n_onion//1_veg
S tilburg.fix                  onion-n_onion//1_veg
A upc-ehu-su.fix               onion-n_onion//1_veg / 9
                               onion-n_onion//2_plant / 1.91786
A upc-ehu-un.fix               onion-n_onion//1_veg / 4
                               onion-n_onion//2_plant / 1.91786
A xeroxceli.fix                onion-n_onion//2_plant
S commonest                    onion-n_onion//1_veg
```

onion 700007:

Heat the oil in a heavy-bottomed pan and add the beef.

Fry, turning frequently to seal the meat.

Add the <tag "onion-n_onion//1_veg"> onion, garlic, carrot, celery and leek and cook for 2 minutes.

```
O avignon.fix                  onion-n_onion//1_veg / 0.97
A clres.fix                    onion-n_onion//1_veg
S durham.fix                   onion-n_onion//1_veg
S ets-pu.fix                   onion-n_onion//1_veg
S grling-sdm.fix               onion-n_onion//1_veg
O hopkins.fix                  onion-n_onion//1_veg
S korea.fix                    onion-n_onion//1_veg
A malaysia.fix                 onion-n_onion//1_veg
A manitoba.dl.dictonly.fix     onion-n_onion//1_veg / 0.46972
A manitoba.dl.fix              onion-n_onion//1_veg / 0.575217
S manitoba.ks.fix              onion-n_onion//1_veg
```

```
A suss.fix                    onion-n_onion//1_veg
S tilburg.fix                 onion-n_onion//1_veg
A upc-ehu-su.fix              onion-n_onion//1_veg / 8
                              onion-n_onion//2_plant / 3.15217
A upc-ehu-un.fix              onion-n_onion//1_veg / 3
                              onion-n_onion//2_plant / 3.15217
A xeroxceli.fix               onion-n_onion//1_veg
S commonest                   onion-n_onion//1_veg
```

onion 700008:

Pre-heat the oven to gas mark 1 " / " 2 60&degree. 1 " / " 2 25&degree.F. 2, Heat the oil and butter together in a heavy pan or casserole dish, add the <tag "onion-n_onion//1_veg"> onion and peppers and cook until soft.

```
O avignon.fix                 onion-n_onion//1_veg / 0.85
                              onion-n_spring_onion_spring / 0.12
A clres.fix                   onion-n_onion//1_veg
S durham.fix                  onion-n_onion//1_veg
S ets-pu.fix                  onion-n_onion//1_veg
S grling-sdm.fix              onion-n_onion//1_veg
O hopkins.fix                 onion-n_onion//1_veg
S korea.fix                   onion-n_onion//1_veg
A manitoba.dl.dictonly.fix    onion-n_onion//1_veg / 0.338419
A manitoba.dl.fix             onion-n_onion//1_veg / 0.438757
S manitoba.ks.fix             onion-n_onion//1_veg
A ottawa.ret.fix              onion-n_onion//2_plant
A suss.fix                    onion-n_onion//1_veg
O sussex.fix                  onion-n_onion//1_veg
S tilburg.fix                 onion-n_onion//1_veg
A upc-ehu-su.fix              onion-n_onion//1_veg / 9
                              onion-n_onion//2_plant / 1.85
A upc-ehu-un.fix              onion-n_onion//1_veg / 4
                              onion-n_onion//2_plant / 1.85
A xeroxceli.fix               onion-n_onion//1_veg
S commonest                   onion-n_onion//1_veg
```

onion 700009:

If you have no greenhouse then sow one row thinly and transplant the thinnings, raking in two handfuls of fertiliser per square yard before sowing or planting.

Spring <tag "onion-n_spring_onion_spring"> onions are treated in the same way as radish, while parsnips must go in early, should be sown in shallow drills with around three or four seeds together at six inch intervals after a handful of fertiliser per square yard has been worked in.

```
O avignon.fix                 onion-n_onion//1_veg / 0.62
                              onion-n_onion//2_plant / 0.35
A clres.fix                   onion-n_onion//1_veg
S durham.fix                  onion-n_spring_onion_spring
S ets-pu.fix                  onion-n_onion//1_veg
S grling-sdm.fix              onion-n_onion//1_veg
O hopkins.fix                 onion-n_spring_onion_spring
S korea.fix                   onion-n_onion//2_plant
A manitoba.dl.dictonly.fix    onion-n_onion//1_veg / 0.300157
A manitoba.dl.fix             onion-n_onion//1_veg / 0.346033
S manitoba.ks.fix             onion-n_onion//1_veg
A suss.fix                    onion-n_spring_onion_spring
S tilburg.fix                 onion-n_onion//1_veg
A upc-ehu-su.fix              onion-n_onion//1_veg / 4
                              onion-n_spring_onion_spring / 11
                              onion-n_onion//2_plant / 1.91923
```

```
A upc-ehu-un.fix              onion-n_onion//1_veg / 4
                              onion-n_spring_onion_spring / 7
                              onion-n_onion//2_plant / 1.91923
A xeroxceli.fix               onion-n_onion//2_plant
S commonest                   onion-n_onion//1_veg
```

onion 700010:

One of the best bulbous plants for drying is Allium albopilosum (christophii).

This ornamental <tag "onion-n_onion//2_plant"> onion blooms in June with large globe-shaped flowers up to ten inches in diameter, with small star-shaped silver-lilac flowers.

```
O avignon.fix                 onion-n_onion//1_veg / 0.76
A clres.fix                   onion-n_onion//1_veg
S durham.fix                  onion-n_onion//1_veg
S ets-pu.fix                  onion-n_onion//2_plant
S grling-sdm.fix              onion-n_onion//1_veg
O hopkins.fix                 onion-n_onion//1_veg
S korea.fix                   onion-n_onion//1_veg
A manitoba.dl.dictonly.fix    onion-n_onion//1_veg / 0.301419
A manitoba.dl.fix             onion-n_onion//1_veg / 0.354529
S manitoba.ks.fix             onion-n_onion//1_veg
A suss.fix                    onion-n_onion//1_veg
S tilburg.fix                 onion-n_onion_dome_basil
A upc-ehu-su.fix              onion-n_onion//1_veg / 8
                              onion-n_onion//2_plant / 2.62727
A upc-ehu-un.fix              onion-n_onion//1_veg / 3
                              onion-n_onion//2_plant / 2.62727
S commonest                   onion-n_onion//1_veg
```

onion 700011:

Marinade:

2-3 cloves garlic, crushed 1 tsp ground cumin 1 tsp ground cinnamon 1 / 2 tsp ground coriander 1 / 2 tsp paprika 2-3tbs olive oil Juice of 1-2 lemons Pinch cayenne pepper Salt and freshly-ground pepper 1 1 / 2 lb cod cheeks, skinned 8 dates, stoned and halved, 4 young turnips, peeled and thinly sliced 1 / 2 lb blanched green beans, sliced 1 / 2 lb <tag "onion-n_onion//1_veg"> onions, sliced Bunch parsley Preparation: Thoroughly mix all marinade ingredients: leave fish in the mixture for at least one hour, and up to five hours.

```
O avignon.fix                 onion-n_onion//1_veg / 0.75
A clres.fix                   -1
S durham.fix                  onion-n_onion//1_veg
S ets-pu.fix                  onion-n_onion//1_veg
S grling-sdm.fix              onion-n_onion//1_veg
O hopkins.fix                 onion-n_onion//1_veg
S korea.fix                   onion-n_onion//1_veg
A manitoba.dl.dictonly.fix    onion-n_onion//1_veg / 0.223469
A manitoba.dl.fix             onion-n_onion//1_veg / 0.225629
S manitoba.ks.fix             onion-n_onion//1_veg
A ottawa.ret.fix              onion-n_onion//1_veg
A suss.fix                    onion-n_onion//1_veg
S tilburg.fix                 onion-n_onion//1_veg
A upc-ehu-su.fix              onion-n_onion//1_veg / 10
                              onion-n_onion//2_plant / 2.23333
A upc-ehu-un.fix              onion-n_onion//1_veg / 5
                              onion-n_onion//2_plant / 2.23333
A xeroxceli.fix               onion-n_onion//2_plant
S commonest                   onion-n_onion//1_veg
```

generous 700002:

As he said in another context, "it was a yell rather than a thought."

The wildness of the suggestion that their own father should wait until they had grown up before being allowed access to his own sons revealed, as well as pain, a <tag "generous-a_generous//3_kind"> generous love.

A	clres.fix	generous-a_generous//1_unstint
S	durham.fix	generous-a_generous//1_unstint
S	ets-pu.fix	generous-a_generous//1_unstint
S	grling-sdm.fix	generous-a_generous//2_bigbucks
O	hopkins.fix	generous-a_generous//2_bigbucks
S	korea.fix	generous-a_generous//1_unstint
A	malaysia.fix	generous-a_generous//4_liberal
A	manitoba.dl.dictonly.fix	generous-a_generous//5_copious / 0.428386
A	manitoba.dl.fix	generous-a_generous//2_bigbucks / 0.523471
S	manitoba.ks.fix	generous-a_generous//2_bigbucks
A	suss.fix	generous-a_generous//1_unstint
S	tilburg.fix	generous-a_generous//1_unstint
S	commonest	generous-a_generous//2_bigbucks
S	commonest.subsumer	generous-a_generous//2_bigbucks
S	commonest.trainingonly	generous-a_generous//1_unstint
S	commonest.trainingonly.subsumer	generous-a_generous//1_unstint
S	commonest.trainingonly.main	generous-a_generous//2_bigbucks

generous 700003:

Broderick launches into his reply like a trouper.

"Oh, it was wonderful, fascinating, a rich experience.

He's a very <tag "generous-a_generous//1_unstint or generous-a_generous//3_kind"> generous actor and obviously he's very full."

A	clres.fix	generous-a_generous//1_unstint
S	durham.fix	generous-a_generous//1_unstint
S	ets-pu.fix	generous-a_generous//1_unstint
S	grling-sdm.fix	generous-a_generous//1_unstint
O	hopkins.fix	generous-a_generous//1_unstint
S	korea.fix	generous-a_generous//1_unstint
A	malaysia.fix	generous-a_generous//6_spacious
A	manitoba.dl.dictonly.fix	generous-a_generous//5_copious / 0.425237
A	manitoba.dl.fix	generous-a_generous//1_unstint / 0.591059
S	manitoba.ks.fix	generous-a_generous//1_unstint
A	suss.fix	generous-a_generous//1_unstint
S	tilburg.fix	generous-a_generous//1_unstint
A	xeroxceli.fix	generous-a_generous//1_unstint
S	commonest	generous-a_generous//2_bigbucks
S	commonest.subsumer	generous-a_generous//2_bigbucks
S	commonest.trainingonly	generous-a_generous//1_unstint
S	commonest.trainingonly.subsumer	generous-a_generous//1_unstint
S	commonest.trainingonly.main	generous-a_generous//2_bigbucks

generous 700004:

Man Ray, born Emmanuel Radnitzky of Jewish immigrants in Philadelphia in 1890, renounced deep family and ethnic ties in his allegiance to the cult of absolute artistic freedom.

Paradoxically, his fame as the almost hypnotic photo-portrayer of the leading artistic figures around him, his novel solarisations,

rayographs and cliches de verre (the last two cameraless manipulations of light and chemistry alone), and his original work for Vogue and Harper's became a diamond-studded albatross about the neck of a man who wanted to be recognised, first and foremost, as a painter.

A more <tag "generous-a_generous//5_copious"> generous supply of illustrations might have helped the reader place him in the history of 20th-century art.

A	clres.fix	generous-a_generous//1_unstint
S	durham.fix	generous-a_generous//2_bigbucks
S	ets-pu.fix	generous-a_generous//3_kind
S	grling-sdm.fix	generous-a_generous//5_copious
O	hopkins.fix	generous-a_generous//3_kind
S	korea.fix	generous-a_generous//5_copious
A	malaysia.fix	generous-a_generous//6_spacious
A	manitoba.dl.dictonly.fix	generous-a_generous//2_bigbucks / 0.568391
A	manitoba.dl.fix	generous-a_generous//1_unstint / 0.569436
S	manitoba.ks.fix	generous-a_generous//1_unstint
A	suss.fix	generous-a_generous//1_unstint
S	tilburg.fix	generous-a_generous//5_copious
A	xeroxceli.fix	generous-a_generous//5_copious
S	commonest	generous-a_generous//2_bigbucks
S	commonest.subsumer	generous-a_generous//2_bigbucks
S	commonest.trainingonly	generous-a_generous//1_unstint
S	commonest.trainingonly.subsumer	generous-a_generous//1_unstint
S	commonest.trainingonly.main	generous-a_generous//2_bigbucks

generous 700005:

Mrs Brown said: "It's a really great way of attracting people's attention, because they can't fail to notice us."

"People have been very <tag "generous-a_generous//1_unstint"> generous and we raised about #200 within the first few hours."

A	clres.fix	generous-a_generous//1_unstint
S	durham.fix	generous-a_generous//1_unstint
S	ets-pu.fix	generous-a_generous//1_unstint
S	grling-sdm.fix	generous-a_generous//1_unstint
O	hopkins.fix	generous-a_generous//1_unstint
S	korea.fix	generous-a_generous//1_unstint
A	malaysia.fix	generous-a_generous//6_spacious
A	manitoba.dl.dictonly.fix	generous-a_generous//3_kind / 0.551274
A	manitoba.dl.fix	generous-a_generous//1_unstint / 0.688915
S	manitoba.ks.fix	generous-a_generous//2_bigbucks
A	suss.fix	generous-a_generous//1_unstint
S	tilburg.fix	generous-a_generous//1_unstint
A	xeroxceli.fix	generous-a_generous//1_unstint
S	commonest	generous-a_generous//2_bigbucks
S	commonest.subsumer	generous-a_generous//2_bigbucks
S	commonest.trainingonly	generous-a_generous//1_unstint
S	commonest.trainingonly.subsumer	generous-a_generous//1_unstint
S	commonest.trainingonly.main	generous-a_generous//2_bigbucks

generous 700006:

A super year for all cash, career and personal affairs.

ARIES (Mar 21-Apr 20): There are some hefty hints being thrown around on Tuesday from folk who may be angling for a favour, a promise or a <tag "generous-a_generous//1_unstint or generous-a_generous//3_kind"> generous gesture.

A	clres.fix	generous-a_generous//1_unstint

```
S durham.fix                        generous-a_generous//4_liberal
S ets-pu.fix                        generous-a_generous//3_kind
S grling-sdm.fix                    generous-a_generous//1_unstint
O hopkins.fix                       generous-a_generous//3_kind
S korea.fix                         generous-a_generous//3_kind
A malaysia.fix                      generous-a_generous//3_kind
A manitoba.dl.dictonly.fix          generous-a_generous//5_copious / 0.410656
A manitoba.dl.fix                   generous-a_generous//3_kind / 0.768432
S manitoba.ks.fix                   generous-a_generous//3_kind
A suss.fix                          generous-a_generous//1_unstint
S tilburg.fix                       generous-a_generous//3_kind
A xeroxceli.fix                     generous-a_generous//2_bigbucks
S commonest                         generous-a_generous//2_bigbucks
S commonest.subsumer                generous-a_generous//2_bigbucks
S commonest.trainingonly            generous-a_generous//1_unstint
S commonest.trainingonly.subsumer generous-a_generous//1_unstint
S commonest.trainingonly.main       generous-a_generous//2_bigbucks
```

generous 700007:

Seconds later, airborne missiles whooshed through the air from all directions, apparently aimed at our heads.

It would be <tag "generous-a_generous//3_kind or generous-a_generous//4_liberal"> generous to call them fireworks, but that implies something decorative, to which one's response is `Aaah", not `Aaagh".

```
A clres.fix                         generous-a_generous//1_unstint
S durham.fix                        generous-a_generous//3_kind
S ets-pu.fix                        generous-a_generous//1_unstint
S grling-sdm.fix                    generous-a_generous//1_unstint
O hopkins.fix                       generous-a_generous//1_unstint
S korea.fix                         generous-a_generous//3_kind
A malaysia.fix                      generous-a_generous//3_kind
A manitoba.dl.dictonly.fix          generous-a_generous//1_unstint / 0.598882
A manitoba.dl.fix                   generous-a_generous//1_unstint / 0.84501
S manitoba.ks.fix                   generous-a_generous//1_unstint
A suss.fix                          generous-a_generous//1_unstint
S tilburg.fix                       generous-a_generous//1_unstint
A xeroxceli.fix                     generous-a_generous//1_unstint
S commonest                         generous-a_generous//2_bigbucks
S commonest.subsumer                generous-a_generous//2_bigbucks
S commonest.trainingonly            generous-a_generous//1_unstint
S commonest.trainingonly.subsumer generous-a_generous//1_unstint
S commonest.trainingonly.main       generous-a_generous//2_bigbucks
```

generous 700008:

Although he has spent most of his working life in academia he did have an eight-year stint, from 1963, in industrial research.

Industry is <tag "generous-a_generous//1_unstint"> generous to Imperial &dash. it endows chairs, sponsors students and gives the college millions of pounds of research contracts every year &dash. but, despite that, Ash is still very critical of it.

```
A clres.fix                         generous-a_generous//1_unstint
S durham.fix                        generous-a_generous//1_unstint
S ets-pu.fix                        generous-a_generous//1_unstint
S grling-sdm.fix                    generous-a_generous//1_unstint
O hopkins.fix                       generous-a_generous//2_bigbucks
S korea.fix                         generous-a_generous//1_unstint
A malaysia.fix                      generous-a_generous//1_unstint
A manitoba.dl.dictonly.fix          generous-a_generous//3_kind / 0.440324
A manitoba.dl.fix                   generous-a_generous//3_kind / 0.501225
```

```
S manitoba.ks.fix                   generous-a_generous//2_bigbucks
A suss.fix                          generous-a_generous//1_unstint
S tilburg.fix                       generous-a_generous//1_unstint
A xeroxceli.fix                     generous-a_generous//1_unstint
S commonest                         generous-a_generous//2_bigbucks
S commonest.subsumer                generous-a_generous//2_bigbucks
S commonest.trainingonly            generous-a_generous//1_unstint
S commonest.trainingonly.subsumer generous-a_generous//1_unstint
S commonest.trainingonly.main       generous-a_generous//2_bigbucks
```

generous 700009:

This was typical of the constant negotiation and compromise that characterised the wars.

The Dunstanburgh agreement was made at Christmas-time in 1462, but it was not just the season which put the Yorkist government in a <tag "generous-a_generous//3_kind"> generous mood.

```
A clres.fix                         generous-a_generous//1_unstint
S durham.fix                        generous-a_generous//2_bigbucks
S ets-pu.fix                        generous-a_generous//1_unstint
S grling-sdm.fix                    generous-a_generous//5_copious
O hopkins.fix                       generous-a_generous//1_unstint
S korea.fix                         generous-a_generous//3_kind
A malaysia.fix                      generous-a_generous//3_kind
A manitoba.dl.dictonly.fix          generous-a_generous//1_unstint / 0.497144
A manitoba.dl.fix                   generous-a_generous//2_bigbucks / 0.632514
S manitoba.ks.fix                   generous-a_generous//2_bigbucks
A suss.fix                          generous-a_generous//1_unstint
S tilburg.fix                       generous-a_generous//6_spacious
A xeroxceli.fix                     generous-a_generous//3_kind
S commonest                         generous-a_generous//2_bigbucks
S commonest.subsumer                generous-a_generous//2_bigbucks
S commonest.trainingonly            generous-a_generous//1_unstint
S commonest.trainingonly.subsumer generous-a_generous//1_unstint
S commonest.trainingonly.main       generous-a_generous//2_bigbucks
```

generous 700010:

The third concert, of Brahms's Third and First symphonies, revealed the new Karajan at his most lovable, for these were natural, emotional, and &dash. let the word escape at last &dash. profound interpretations: voyages of discovery; loving traversals of familiar, exciting ground with a fresh eye and mind, in the company of someone prepared to linger here, to exclaim there; summations towards which many of his earlier, less intimate performances of the works had led.

Karajan had pitched camp with Legge and the Philharmonia in 1949 when a <tag "generous-a_generous//1_unstint or generous-a_generous//2_bigbucks"> generous grant from the Maharaja of Mysore had stabilized the orchestra's finances and opened up the possibility, in collaboration with EMI, of extensive recording, not only of the classic repertory but of works that caught Karajan's and Legge's fancy: Balakirev's First Symphony, Roussel's Fourth Symphony, the still formidably difficult Music for Strings, Percussion, and Celesta by Barto&acute.k, and some English music, too.

```
A clres.fix                         generous-a_generous//1_unstint
S durham.fix                        generous-a_generous//5_copious
S ets-pu.fix                        generous-a_generous//1_unstint
S grling-sdm.fix                    generous-a_generous//3_kind
O hopkins.fix                       generous-a_generous//2_bigbucks
S korea.fix                         generous-a_generous//2_bigbucks
A malaysia.fix                      generous-a_generous//3_kind
```

```
A manitoba.dl.dictonly.fix           generous-a_generous//2_bigbucks / 0.461805
A manitoba.dl.fix                    generous-a_generous//2_bigbucks / 0.542222
S manitoba.ks.fix                    generous-a_generous//1_unstint
A suss.fix                           generous-a_generous//1_unstint
S tilburg.fix                        generous-a_generous//2_bigbucks
A xeroxceli.fix                      generous-a_generous//1_unstint
S commonest                          generous-a_generous//2_bigbucks
S commonest.subsumer                 generous-a_generous//2_bigbucks
S commonest.trainingonly             generous-a_generous//1_unstint
S commonest.trainingonly.subsumer    generous-a_generous//1_unstint
S commonest.trainingonly.main        generous-a_generous//2_bigbucks
```

SENSEVAL/ROMANSEVAL

the Italian Systems:

a few observations

Nicoletta Calzolari
ILC - Pisa

Number of Senses

Incidence of polysemy?

POS	2	3	4	5	6	7	8	9	10	11	16
NOUNS	2	4	3	4	2	1	2	1			
ADJECTIVES	3	2	6	2	2	1	2	1			
VERBS	5	4	3	1	1	2	1	1		1	1
TOTAL	10	10	12	7	5	4	5	3		1	1

Polysemy & Performance

- no clear correlation between polysemy & performance of systems, e.g.
  - *alto:* 8 senses, wrong=14, right=34(12(full)+22(part))
  - *biologico:* 3 senses, wrong=11, right= 27(1(full)+26(part))
  - *breve:* 4 senses, wrong=10, right= 41(26(full)+15(part))
  - *chiaro:* 9 senses, wrong=20, right= 26(7(full)+19(part)),?=5 (3 multiple only)
  - *civile:* 5 senses, wrong= 8, right= 40(12(full)+28(part)),?=3 (15 multiple)
  - *eccezionale:* 2 senses, wrong= 8, right= 22(16(full)+6(part))

Adjectives: Polysemy & Performance

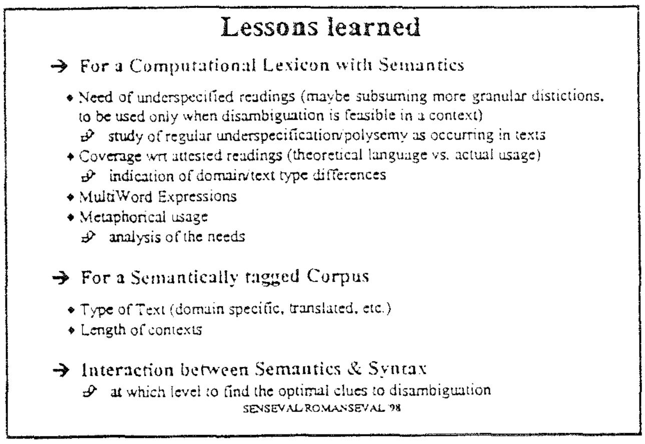
	Senses	Wrong %	Right %	Fully %	Partial %	? %	Multiple Tag
alto	8(5)	29.1	70.9				18
biologico	3	29.9	71.1	2.6	68.4		20
breve	4	19.6	80.4	50.9	29.4		13
chiaro	9(4)	39.2	50.9	13.7	37.2	9.8	3
civile	5	15.6	78.4	23.5	54.9	5.8	15
eccezionale	2	26.7	73.3	53.3	20		1
legale						3.9	
libero							3
nuovo	7	60.7					0
particolare	2	17.6					
pieno	6						
popolare	4	7.1					

Nouns: Polysemy & Performance

	Senses	Wrong %	Right %	Fully %	Partial %	? %	Multiple
agente	3	-	98.1	94.1	3.9	1.9	2
campagna	4	5.9	94.1	84.3	9.8		0
capo	7	3.9	90.1	27.4	62.7	5.9	30
centro	[illegible]		86.2	74.5	11.7	7.84	1
compagnia	6						
comunicazione	5						
concentrazione	[illegible]						0
condizione	4						
corso	8						
costituzione	3						
detenzione	2						
lancio	3	50	50				0

Verbs: Difference in Performance

	Senses	Contexts	Wrong	?	Multiple	Wrong	?	Multiple
arrivare	3	29	1	12	1	8		
		25	0	9	0	0		
chiedere	4	51	3	5	5	0		51
					4			41
		36	=	=	0	=		36
comprendere	2	51	13	15		8		
concludere	3	51	2	12		17		
		42	=	6		16		
mantenere	5	51	10	13		0		
		34	5	8		=		
		26	=	4		=		



Lessons learned

→ For a Computational Lexicon with Semantics

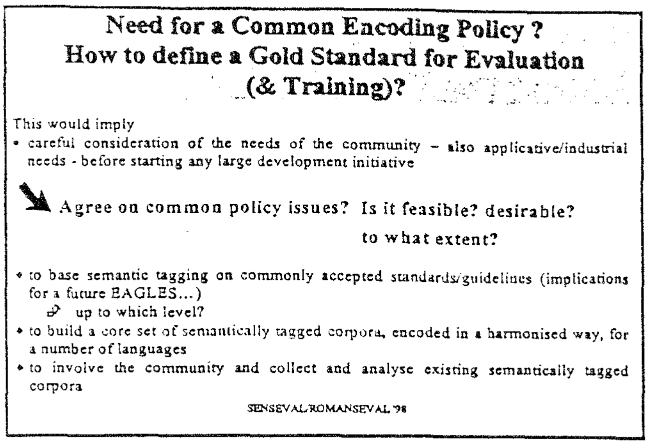
- Need of underspecified readings (maybe subsuming more granular distictions, to be used only when disambiguation is feasible in a context)
  - study of regular underspecification/polysemy as occurring in texts
- Coverage wrt attested readings (theoretical language vs. actual usage)
  - indication of domain/text type differences
- MultiWord Expressions
- Metaphorical usage
  - analysis of the needs

→ For a Semantically tagged Corpus

- Type of Text (domain specific, translated, etc.)
- Length of contexts

→ Interaction between Semantics & Syntax

- at which level to find the optimal clues to disambiguation



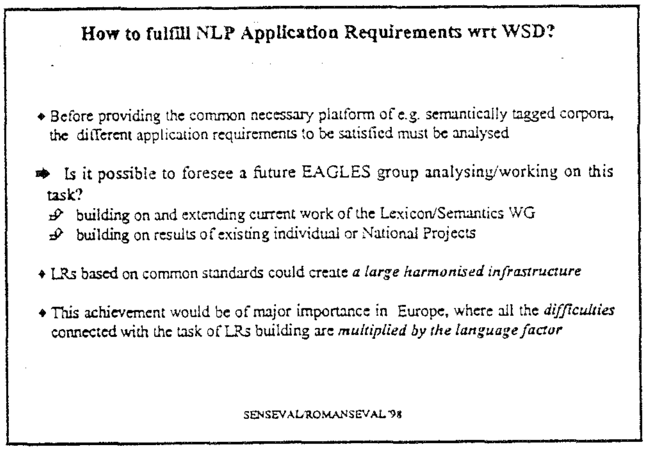
Need for a Common Encoding Policy ? How to define a Gold Standard for Evaluation (& Training)?

This would imply

- careful consideration of the needs of the community – also applicative/industrial needs - before starting any large development initiative

➘ Agree on common policy issues? Is it feasible? desirable? to what extent?

- to base semantic tagging on commonly accepted standards/guidelines (implications for a future EAGLES...)
  - up to which level?
- to build a core set of semantically tagged corpora, encoded in a harmonised way, for a number of languages
- to involve the community and collect and analyse existing semantically tagged corpora



How to fulfill NLP Application Requirements wrt WSD?

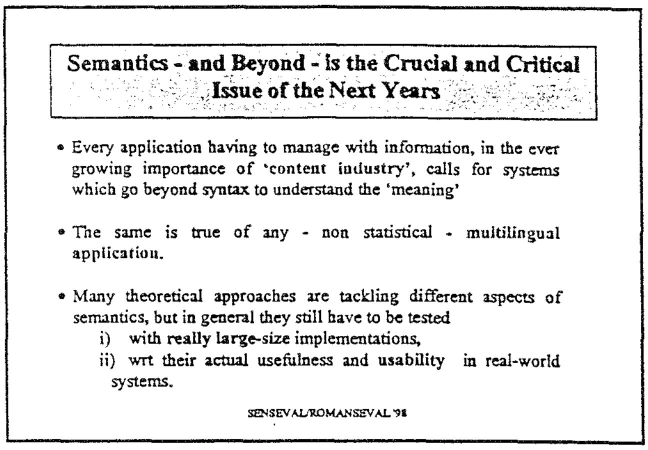
- Before providing the common necessary platform of e.g. semantically tagged corpora, the different application requirements to be satisfied must be analysed

➡ Is it possible to foresee a future EAGLES group analysing/working on this task?

- building on and extending current work of the Lexicon/Semantics WG
- building on results of existing individual or National Projects

- LRs based on common standards could create *a large harmonised infrastructure*

- This achievement would be of major importance in Europe, where all the *difficulties* connected with the task of LRs building are *multiplied by the language factor*

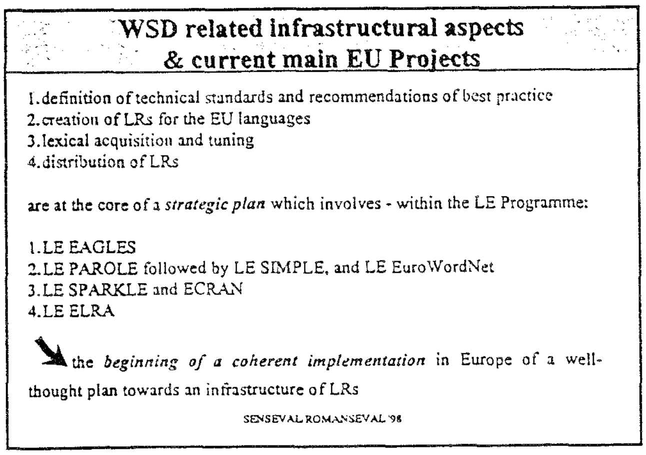


Semantics - and Beyond - is the Crucial and Critical Issue of the Next Years

- Every application having to manage with information, in the ever growing importance of 'content industry', calls for systems which go beyond syntax to understand the 'meaning'

- The same is true of any - non statistical - multilingual application.

- Many theoretical approaches are tackling different aspects of semantics, but in general they still have to be tested
  i) with really large-size implementations,
  ii) wrt their actual usefulness and usability in real-world systems.



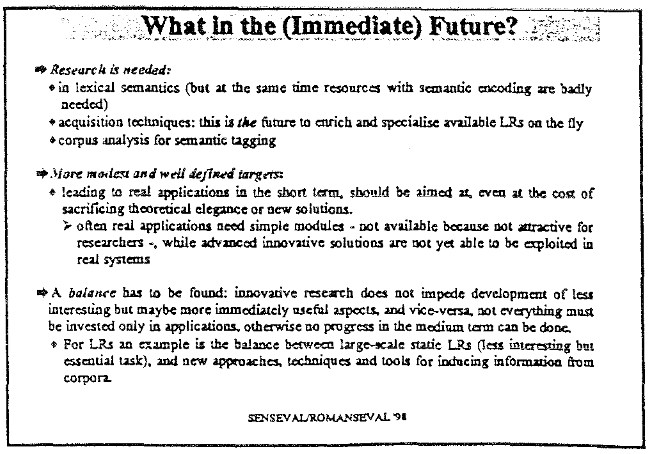
WSD related infrastructural aspects & current main EU Projects

1.definition of technical standards and recommendations of best practice
2.creation of LRs for the EU languages
3.lexical acquisition and tuning
4.distribution of LRs

are at the core of a *strategic plan* which involves - within the LE Programme:

1.LE EAGLES
2.LE PAROLE followed by LE SIMPLE, and LE EuroWordNet
3.LE SPARKLE and ECRAN
4.LE ELRA

➘ the *beginning of a coherent implementation* in Europe of a well-thought plan towards an infrastructure of LRs



What in the (Immediate) Future?

➡ *Research is needed:*

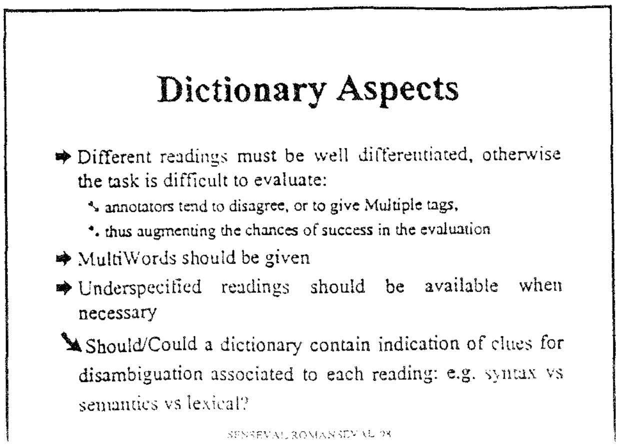
- in lexical semantics (but at the same time resources with semantic encoding are badly needed)
- acquisition techniques: this is *the* future to enrich and specialise available LRs on the fly
- corpus analysis for semantic tagging

➡ *More modest and well defined targets:*

- leading to real applications in the short term, should be aimed at, even at the cost of sacrificing theoretical elegance or new solutions.
  - ➢ often real applications need simple modules - not available because not attractive for researchers -, while advanced innovative solutions are not yet able to be exploited in real systems

➡ A *balance* has to be found: innovative research does not impede development of less interesting but maybe more immediately useful aspects, and vice-versa, not everything must be invested only in applications, otherwise no progress in the medium term can be done.

- For LRs an example is the balance between large-scale static LRs (less interesting but essential task), and new approaches, techniques and tools for inducing information from corpora.



Dictionary Aspects

➡ Different readings must be well differentiated, otherwise the task is difficult to evaluate:

- annotators tend to disagree, or to give Multiple tags,
- thus augmenting the chances of success in the evaluation

➡ MultiWords should be given

➡ Underspecified readings should be available when necessary

➘ Should/Could a dictionary contain indication of clues for disambiguation associated to each reading: e.g. syntax vs semantics vs lexical?

Selecting decomposable models for word-sense disambiguation: The *grling-sdm* system

Tom O'Hara[1], Janyce Wiebe[1], and Rebecca Bruce[2]

[1]Department of Computer Science &
Computing Research Laboratory
New Mexico State University

[2]Department of Computer Science
University of North Carolina at Asheville

Methodology

- Probabilistic classification
- Supervised approach
- Model search
- Collocational feature organizations

Feature description

- "Knowledge-lite" approach
- Shallow linguistic features
  - part-of-speech for immediate context
  - unconstrained collocations for each sense

example:

However, salad crops such as lettuce$_{NN}$ and$_{CC}$ <tag "528344">onions$_{NNS}$</tag> are$_{VB}$ always$_{RB}$ popular, while those like broad beans, peas and spinach are ...

<{NN, CC, NNS, VB, RB, 0, 1, 0}, 528344>

Results for supervised systems

Recall

Task	Mean	Stdev	grling-sdm
onion-n	.735	.232	.846
generous-a	.462	.127	.476
shake-p	.598	.140	.596

Precision

Task	Mean	Stdev	grling-sdm
onion-n	.857	.035	.846
generous-a	.520	.045	.482
shake-p	.667	.061	.644

Note: results over fine-grained scores

Improvement over Naive Bayes

Word	Entropy	Baseline	Naive Bayes	Model Selected	Gain
sick	2.969	30.8	56.8	65.1	8.3
curious	0.833	76.9	83.0	87.8	4.8
beam	2.950	35.4	61.1	65.8	4.8
brick	2.289	47.9	68.1	71.7	3.6
drain	3.253	19.3	57.3	60.9	3.6
bake	2.691	23.8	79.1	80.9	1.8

notes: dry-run data; 10-fold cross-validation; statistically significant ($p<0.05$)

Conclusion

- “Knowledge-lite” approach to WSD
- Focus on methodology
- Thanks to:
  Ted Pedersen
  SENSEVAL coordinators
  Oxford University Press & other sponsors

Ronan Pichon, Pascale Sébillot
IRISA, Campus de Beaulieu, 35042 Rennes cedex, France
tel: 33 2 99 84 74 50 / 73 17; fax: 33 2 99 84 71 71
email: rpichon@irisa.fr, sebillot@irisa.fr

The description of the method, which is here given for adjectives, is identical for verbs and nouns.

Starting point:

Each occurrence of an adjective is associated with its own lemma, all the nouns of the sentence where it occurs, all the verbs of the sentence where it occurs, and all the other adjectives of the sentence where it occurs.

Here is a more precise description of the associated information.

First of all, the elements of the corpus are coded in the following way:

Each (different) noun gets a number, between 1 and the cardinal of the set of the different nouns in the corpus (that is 5237).

Each (different) verb gets a number, between 1 and the cardinal of the set of the different verbs in the corpus (that is 1915).

Each (different) adjective gets a number, between 1 and the cardinal of the set of the different adjectives in the corpus (that is 2217).

Therefore, at the beginning of the treatment, each occurrence of an adjective gets a vector of attributes, which consists of 4 vectors:

- 1 vector of its own lemma; for example 10:1, if it corresponds to the adjective number 10;
- 1 vector of the nouns which occur in the same sentence; for example 1:1, 5:2, 56:1, if the nouns 1, 5 and 56 respectively appear once, twice and once in this sentence;
- 1 vector of the verbs which occur in the same sentence; for example 2:1, 4:1, if the verbs 2 and 4 both appear once in this sentence;
- 1 vector of the adjectives which occur in the same sentence; for example 8:1, 25:1, if the adjectives 8 and 25 both appear once in this sentence.

This is the data structure for each occurrence of an adjective, but it can also be considered as that of a cluster of adjectives after several steps of association, whose method is now described.

The clustering method:

In order to improve the speed of the method, the set of the occurrences of adjectives is cut into subsets of 1000 elements, that are treated separately, until a 10% reduction of their sizes.

For each cluster (at the beginning, of one adjective, then of several adjectives), we calculate an association coefficient with every other cluster (that is, with the 999 other clusters, for the first time). During the calculation of the coefficients, the 50 hightest values of association coefficients are memorized; of course, if the association between cluster C_i and cluster C_j is already selected (C_j is therefore the cluster which is the most strongly associated with C_i, and conversely), the associations following the frames (C_i,x) or (C_j,y) cannot be kept among the 49 other strongest associations. At the end of the calculus, the 50 cluster links that have been determined as the strongest at this step are aggregated (sum of the corresponding vectors). Then a new step begins. When all the initial subsets of 1000 clusters are reduced to 900 elements (10% reduction), all the remaining clusters are put together and the clustering method is re-applied, till the obtaining of 1000 clusters (arbitrarily fixed value).

Calculus of the association coefficient between two clusters:

$$\frac{\mathtt{Li}}{\mathtt{||Li||}} \cdot \frac{\mathtt{Lj}}{\mathtt{||Lj||}} + \frac{\mathtt{Ni}}{\mathtt{||Ni||}} \cdot \frac{\mathtt{Nj}}{\mathtt{||Nj||}} + \frac{\mathtt{Vi}}{\mathtt{||Vi||}} \cdot \frac{\mathtt{Vj}}{\mathtt{||Vj||}} + \frac{\mathtt{Ai}}{\mathtt{||Ai||}} \cdot \frac{\mathtt{Aj}}{\mathtt{||Aj||}}$$

.: scalar product (between normalized vectors, so that a vector with a high number of elements has no higher weight than others)

L: vector of lemmas, N: vector of nouns, V: vector of verbs, A: vector of adjectives.

Therefore, the association coefficient value is between 0 and 4 (for example, 4 corresponds to 2 occurrences of a same adjective found in two identical sentences).

Some Problems - Some Solutions:

Concerning verbs, results are not good. In fact, we have stopped the search of the meanings of the test occurrences. One explanation: there are greedy clusters which "swallow" a lot of verbs; therefore, the interpretation of the class is impossible. This greedy cluster phenomenon also happens for other categories, but it is very accentuated for the verbs. A "normal" class contains about 30-50 elements (that means about 6 to 8 distinct lemmas); a greedy cluster can contain 2000 elements; the maximal cluster for verbs that we have found had 20000 elements.

Different contexts for nouns, verbs and adjectives will probably improve the results. For example, we think that for adjectives, it will be better to consider a closer context (better than the whole sentence).

Selecting decomposable models for word-sense disambiguation: The *grling-sdm* system*

Tom O'Hara and **Janyce Wiebe**
Department of Computer Science and
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
tomohara, wiebe@cs.nmsu.edu

Rebecca Bruce
Department of Computer Science
University of North Carolina at Asheville
Asheville, NC 28804-3299
bruce@cs.unca.edu

Abstract

This paper describes the *grling-sdm* system which is a supervised statistical classifier participating in the 1998 SENSEVAL competition for word-sense disambiguation. *grling-sdm* uses model search to select decomposable models describing the interdependencies among the features describing the data. These types of models have been found to be advantageous in terms of efficiency and graphical representation. Results over the SENSEVAL evaluation data are discussed. In addition, experiments over the dry-run data are included to show how the system performs relative to Naive Bayes classifiers, which are commonly used in natural language processing.

1 Introduction

A probabilistic classifier assigns the most probable sense to a word, based on a probabilistic model of the interdependencies among the word and a set of input features. There are several approaches to determining which models to use. In natural language processing, fixed models are often assumed, but improvements can often be achieved by selecting the model based on characteristics of the data. The *grling-sdm*[1] system was developed at New Mexico State University and the University of North Carolina at Asheville to test the use of probabilistic model selection for word-sense disambiguation in the SENSEVAL competition (Kilgarriff, 1998). This builds upon the approach laid out in (Bruce and Wiebe, 1994) and later refined in (Pedersen and Bruce, 1997) and (Wiebe et al., 1997).

Shallow linguistic features are used in the classification model: part-of-speech in the immediate context and collocations[2] that are indicative of particular senses. Note that the focus of our research has been on the underlying methodology for model formulation and feature representation. One important aspect of this is the investigation of beneficial representations for collocational features.

Manually-annotated training data (*tagged data*) is used to determine the relationships among the features, so this is a supervised learning approach. However, no additional knowledge is incorporated into the system. In particular, the HECTOR definitions and examples are not utilized. Although this "knowledge-lite" approach did not achieve the best results for SENSEVAL, it has performed well on other word-sense disambiguation tasks. In particular, we will show that our approach can lead to significant improvements over Naive Bayes classifiers (i.e., those that make the simplifying assumption of independence among the feature variables given the classification variable). Naive Bayes classifiers have been shown to work remarkably well in many machine learning applications. Therefore, the improvements over them highlight the strengths of this approach.

Supervised approaches to word-sense disambiguation have been shown to achieve high accuracy without the incorporation of domain-specific knowledge (Bruce and Wiebe, 1994; Ng and Lee, 1996). The main drawback is that tagged training data is required, which is often difficult to obtain on a large-scale. Nonetheless, we believe that supervised approaches will continue to play an important role in natural language processing. For example, as outlined in (Ng, 1997), it is feasible to obtain tagged data for the most common polysemous words in a language given a concerted tagging effort.

After presenting a brief overview of statistical classifiers in section 2, we will present an overview of the system in section 3 and then present the results on the tasks chosen for comparison in section 4 (these tasks were selected by the SENSEVAL coordinators). Then to illustrate the strengths of the approach in the context of supervised learning, we present results over the data distributed for the dry-run in section 5.

* This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776.

[1] GraphLing is the name of a project researching graphical models for linguistic applications. SDM refers to supervised decomposable model search.

[2] Collocations are used here in the broader sense of words that co-occur often in context: there are no constraints on word order, etc.

2 Statistical Classification

The goal of statistical classifiers is to predict the value of a classification variable given values for variables describing the input or *features*. This is done as follows for the simple case of Bayesian Classifiers (Charniak, 1993; Franz, 1996).

Given: Set of *features*, F_i, describing input, I.
Determine the *class* value, C_j, that best fits the input:

1. Collect large sample of known classifications: $< \{f_1, ..., f_n\}, c_i >$
2. Estimate probability of each feature given each class value: $\hat{P}(F_i = f_i|C_j = c_j) = \frac{freq(f_i,\ c_j)}{freq(c_j)}$
3. Choose value maximizing the probability of the class given the features:

$$\begin{aligned}
&\hat{P}(C_j = c_j|F_1 = f_1, ..., F_n = f_n) \\
&= P(f_1, ..., f_n|c_j)P(c_j)/P(f_1, ..., f_n) \\
&= P(f_1|c_j)P(f_2|f_1, c_j)...P(f_n|f_{n-1}, ..., f_1, c_j) \\
&\quad P(c_j)\alpha \\
&= \textstyle\prod_i P(f_i|c_j)P(c_j)\alpha
\end{aligned}$$

where α is a normalizing constant

The first step determines the tagged data that the classifier uses for estimating various parameters of the statistical model. In this case, the variables are assumed to be independent given the value of the classification variable. Therefore, in the second step, the only parameters to be estimated are the conditional probabilities of the feature values given the class value ($P(f_i|c_j)$). The final step successively uses Bayes' Rule, the Chain Rule, and the conditional independence assumption to simplify the calculation of the probability of each class value given the observed features.

Classifiers based on this assumption are called Naive Bayes classifiers. These often perform well in practice because more complex models often suffer from lack of sufficient training data. For example, in a comparative experiment of different machine learning algorithms for word-sense disambiguation using the same features, Mooney (1996) found that Naive Bayes was better than any other method he tried.

3 The *grling-sdm* system

As shown in (Bruce and Wiebe, 1994), it is often advantageous to determine the form of the model (i.e., relationships among the variables), rather than assuming a fixed model as done by Naive Bayes classifiers. The *grling-sdm* system that we developed for SENSEVAL is based on this approach.

Specifically, *grling-sdm* uses model search to select the decomposable model describing the relationships among the features. Decomposable models are a subset of graphical models for which closed-form expressions exist for the model forms. As with other types of graphical models, interdependency relationships can be depicted using graphs (either undirected or directed). See (Bruce and Wiebe, 1996) for further details, including the application of these types of models for word-sense disambiguation.

Feature	Description
POS-2	part-of-speech of second word to the left
POS-1	part-of-speech of word to the left
POS	part-of-speech of word itself (morphology)
POS+1	part-of-speech of word to the right
POS+2	part-of-speech of second word to the right
COLL$_1$	occurrence of collocation for sense 1
...	
COLL$_N$	occurrence of collocation for sense N

Table 1: Features used in *grling-sdm*.

Standard feature sets were used in *grling-sdm*, including parts-of-speech of the words in the immediate context, morphology of the target word, and collocations indicative of each sense. These are summarized in table 1. The collocation variable $coll_i$ for each sense S_i is binary, corresponding to the absence or presence of any word in a set specifically chosen for S_i. A word W is chosen for S_i if $(P(S_i|W) - P(S_i))/P(S_i) \geq 0.2$, that is if the relative percent gain in the conditional probability over the prior probability is 20% or higher. This is a variation on the *per-class, binary organization* discussed by (Wiebe et al., 1998). Note that due to time constraints, we didn't use adjacency-based collocational features, which were found to be beneficial in other work (Pedersen and Bruce, 1998; Ng and Lee, 1996).

The classifier maps the feature values for the context of the word to be disambiguated into a distribution over that word's senses. In probabilistic classification, this distribution is defined by a probability model. Several different models are considered by doing a greedy search through the space of all the probability models. During *forward search*, this proceeds from the model of independence (all features are entirely unrelated) by successively adding dependence constraints until reaching the saturated model (all features are interdependent) or until the termination criteria is reached. *Backward search* proceeds in the opposite direction. Again see (Bruce and Wiebe, 1996) for details.

Instead of selecting a single model, the models are averaged using the Naive Mix (Pedersen and Bruce, 1997), which is a form of smoothing. Higher-complexity models are generally desirable since they better describe the data, but there might not be sufficient training data to cover all the combinations needed for the parameter estimates. To handle this problem, the technique of smoothing factors in mul-

Precision for fine-grained distinctions			
Task	Mean	Stdev	grling-sdm
verb	.605	.118	.640
proper	.674	.130	.693
noun	.774	.097	.710
adj	.669	.090	.672
eval	.669	.096	.676
onion-n	.857	.035	.846
generous-a	.520	.045	.482
shake-p	.667	.061	.644

Recall for fine-grained distinctions			
Task	Mean	Stdev	grling-sdm
verb	.546	.188	.635
proper	.524	.187	.542
noun	.560	.182	.536
adj	.563	.174	.590
eval	.549	.176	.575
onion-n	.735	.232	.846
generous-a	.462	.127	.476
shake-p	.598	.140	.596

Table 2: Overall results for supervised systems

tiple models. This can be viewed as incorporating a default mechanism for cases in which there was insufficient data for the use of the complex model.

The system averages three sets of models: the Naive Bayes model; the final model generated by backward search; and the first k models generated by forward search (for some fixed constant k). There is a strong bias towards the use of simpler models because Naive Bayes and the forward search models are included. However, higher-complexity models are considered because results of the backward search are also included. In future work, we plan to investigate other combinations of these models.

4 Results on evaluation data

The overall results for the performance on fine-grained distinctions by the supervised systems participating in SENSEVAL are shown in table 2. The cases are broken down by task type. Three are for tasks that deal exclusively with a single grammatical category: verb, noun and adjectives. In addition, the type *proper* includes proper-nouns as well as each of the three categories. *eval* is the result for all tasks. This table also includes the performance on the tasks chosen for comparison purposes. As can be seen, the system is roughly performing at an average level. It does better with verbs but worse with nouns.

Feature variable assignments:

A	word-sense
B	POS of the word itself
C	POS 2 words to the left
D	POS 1 word to the left
E	POS 1 word to the right
F	POS 2 words to the right
G	coll_000000
H	coll_528344
I	coll_528347

Models generated during search:
Naive Bayes: AB AC AD AE AF AG AH AI
Backward: ABI ACI ADI AEI AFI AGI AHI
A
AH
AG AH
AG AHI
Forward: AGI AHI

Figure 1: Details on model search for *onion-n*.

The remainder of this section presents detailed results on the three tasks that are being highlighted in the SENSEVAL discussions: *onion-n*, *generous-a*, and *shake-p*. In each case, details on the models used by our system will be given, along with graphical representations of representative cases. Also, confusion matrices are given to show which sense distinctions are problematic for our system.

4.1 onion-n

Figure 1 shows the features for *onion-n* along with the models generated in the search. Recall that the coll_N features are binary features with words indicative of the sense N (using the sense ID's instead of the traditional sense numbers). Note that there are only collocational features for 2 of the 5 possible senses, since 3 cases didn't occur in the training data.

Table 3 shows the confusion matrix for *onion-n*. This indicates the number of times the evaluation key was sense i and the system's guess was sense j. (Note that multiple keys were possible, but that this analysis only considers the first one given.) By comparing the column totals versus the row totals in a confusion matrix, discrepancies can be detected in the responses. Here, the system is always using vegetable sense of "onion" (528347).

One source of these discrepancies was that there was 15 test instances for the sense related to "spring onion" (528348) without any corresponding training data. A similar problem was that the plant sense (528344) only occurred twice in the training data. For supervised to work best, the distribution of senses in the training data should reflect that of the test data.

	Response				
Key	_344	_347	_348	_376	
528344	0	26	0	0	26
528347	0	172	0	0	172
528348	0	15	0	0	15
528376	0	1	0	0	1
	0	214	0	0	214

Table 3: Confusion matrix for *onion-n.*

Feature variable assignments:

A	word-sense
B	POS of the word itself
C	POS 2 words to the left
D	POS 1 word to the left
E	POS 1 word to the right
F	POS 2 words to the right
G	coll_000000
H	coll_512274
I	coll_512275
J	coll_512277
K	coll_512309
L	coll_512310
M	coll_512410

Models generated during search:
Naive Bayes: AB AC AD AE AF AG AH AI
AJ AK AL AM
Backward: ABG ACG ADG AEG AFG AGHJ
AGHL AGI AGK AGM
A
AK
AK AL
AJ AK AL
AI AJ AK AL
AH AI AJ AK AL
AE AH AI AJ AK AL
AB AE AH AI AJ AK AL
AB AE AH AI AJ AK AL AM
AB AE AHJ AI AK AL AM
AB AE AHJ AHL AI AK AM
ABG AE AHJ AHL AI AK AM
Forward: ABG AEG AHJ AHL AI AK AM

Figure 2: Details on model search for *generous-a*.

4.2 generous-a

Figure 2 shows the features for *generous-a* along with the models generated in the search, and figures 8 and 9 show graphical representations of two of the models generated during the search. As can be seen, the backward search model is much more complex than the forward search model. Of interest is the interdependencies between the collocation variables for senses 512274 (unstint), 512277 (kind), and 512310 (copious). The confusion matrix (see table 4) reveals that these cases are not being handled well.

	Response						
Key	_274	_275	_277	_309	_310	_410	
512274	40	1	7	20	16	0	84
512275	3	2	1	4	4	0	14
512277	10	2	6	15	7	0	40
512309	15	3	4	29	5	0	56
512310	6	2	1	4	15	0	28
512410	0	0	0	2	0	0	2
	74	10	19	74	47	0	224

Table 4: Confusion matrix for *generous-a.*

Feature variable assignments:

A	word-sense		
B	POS of the word itself		
C	POS 2 words to the left		
D	POS 1 word to the left		
E	POS 1 word to the right		
F	POS 2 words to the right		
G	coll_000000		
H	coll_504336	W	coll_516391
I	coll_504337	X	coll_516399
J	coll_504338	Y	coll_516494
K	coll_504353	Z	coll_516495
L	coll_504355	a	coll_516517
M	coll_504410	b	coll_516519
N	coll_504412	c	coll_516520
O	coll_504537	d	coll_516551
P	coll_504584	e	coll_516567
Q	coll_504585	f	coll_516605
R	coll_504600	g	coll_516626
S	coll_506816	h	coll_516669
T	coll_516365	i	coll_516708
U	coll_516366	j	coll_516772
V	coll_516390	k	coll_516773

Model considered:
Naive Bayes: AB AC AD AE AF AG AH
AI AJ AK AL AM AN AO AP AQ AR
AS AT AU AV AW AX AY AZ Aa Ab
Ac Ad Ae Af Ag Ah Ai Aj Ak

Figure 3: Details on fixed model for *shake-p*.

4.3 shake-p

For practical reasons, we used Naive Bayes for cases, such as *shake-p*, with more than 25 senses (see figure 3). Running this many features is not infeasible for our approach, but we just ran into time constraints for the competition. As mentioned above, the Naive Bayes model assumes all of the features are independent given the classification variable. See figure 10 for a graphical depiction of the interdependencies. Table 5 shows the confusion matrix for senses of "shake", excluding infrequent cases. This indicates that senses 504338 (move) and 504355 (tremble) are being confused.

5 Results using dry-run data

As mentioned above, our focus in this work was on methodology and feature representation, given a fixed set of knowledge. We will show here that experiments over the dry-run data produced a gain over using a Naive Bayes classifier, a commonly used benchmark that performs remarkably well considering its assumptions (Friedman et al., 1997; Leacock et al., 1993; Mooney, 1996). Note that the method of selecting the testing data was different with the dry-run experiments, because there was no predefined test data. Therefore, the test data was produced by randomly partitioning the dry-run data into 90% training data and 10% test data, using 10-fold cross-validation, which is common practice in machine learning.

We applied the same general method[3] to 34 words randomly selected from a set of 38 words in the SENSEVAL dry-run data (Kilgarriff, 1998). 4 words (or roughly 10%) were set aside to allow a held-out test set for a separate system that required analysis of the dictionary entries. The words were chosen so as to leave approximately 10% of the dry-run corpus instances as test data. Thus, the training data for the experiments during each fold covered roughly 81% of the entire dry-run data.

The results are presented in figure 4. Since 10-fold cross validation was performed for each word, there was a total of 340 experiments. On each fold, a forward search with G^2 as the goodness-of-fit test was performed. In addition, we ensured that Naive Bayes was included as a competitor in each fold. For each fold, evaluation on a single held-out portion of the *training* data was performed to choose the final model. The results of applying this model to the actual test set, averaged over folds, are shown in the column labeled *Model Selection*. The results of applying Naive Bayes exclusively (averaged over folds) are shown in the column labeled *Naive Bayes*.

The same types of features were used in each model (shown earlier in table 1): the part of speech tags one place to the left and right of the ambiguous word; the part of speech tags two places to the left and right of the word; the part of speech tag of the word; and a collocation variable for each sense of the word whose representation is *per-class-binary* as presented in (Wiebe et al., 1997). Again, the variable for each sense S is binary, corresponding to the absence or presence of any word in a set specifically chosen for S. A word W is chosen for S if $P(S|W) \geq 0.5$. (Note that this is different from the method used for the evaluation data, because this analysis was performed prior to deciding on the method to be used for the competition.)

As can be seen in the *Gain* column. the model selection procedure achieves an overall average accuracy that is 1.4 percentage points higher than exclusively using the Naive Bayes classifier. Further, we assessed the statistical significance of the differences in accuracy between the two methods for the individual words, using a paired t-test (Cohen, 1995) with a significance level of 0.05. For six of the words (shown in bold face), the model selection performance is significantly better than the performance of exclusively using Naive Bayes. Further, the model selection procedure is not significantly worse than Naive Bayes for any of the words. Figure 5 shows the top 10 cases both for gains and losses in terms of statistical significance (sorted by p-value, which gives the probability of the improvement occurring by chance).

6 Conclusion

In this paper we illustrated the application of supervised learning techniques to word-sense disambiguation. The performance of the *grling-sdm* system was illustrated using comparative evaluations against other supervised SENSEVAL approaches. In addition, it was shown to give significant improvements over Naive Bayes when applied to experiments over the dry-run data. Such improvements illustrate that the approach is viable.

Acknowledgements

We gratefully acknowledge the contributions to this work by Ted Pedersen. We also wish to thank Oxford University Press for use of the HECTOR data in the SENSEVAL project. Similarly, we thank Cambridge University Press, the UK Engineering and Physical Sciences Research Council, ELRA and Addison Wesley Longman for their contributions to SENSEVAL.

References

Bruce, R., and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 139–146.

Bruce, R., and J. Wiebe. 1996. A Method for Learning Decomposable Models Applied to NLP, Tech. Report MCCS-96-301, Computing Research Laboratory, NMSU.

Charniak, E. 1993. *Statistical Language Learning*, Cambridge, MA: MIT Press.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*, Cambridge, MA: MIT Press.

[3] Here only the best model generated is used rather than taking the average.

Word	Entropy	Baseline	Naive Bayes	Model Selection	Gain
sick	2.969	30.8	56.8	65.1	**+8.3**
storm	2.895	39.6	63.4	71.6	+8.2
drift	2.889	31.7	56.0	63.3	+7.3
curious	0.833	76.9	83.0	87.8	**+4.8**
beam	2.950	35.4	61.1	65.8	**+4.8**
drain	3.253	19.3	57.3	60.9	**+3.6**
brick	2.289	47.9	68.1	71.7	**+3.6**
raider	2.216	36.2	79.6	82.8	+3.3
dawn	2.328	47.0	74.3	77.3	+3.0
sugar	1.786	52.9	82.5	84.9	+2.4
creamy	1.012	68.0	72.3	74.5	+2.3
bake	2.691	23.8	79.1	80.9	**+1.8**
impress	0.758	85.6	89.3	90.8	+1.6
govern	2.139	43.4	67.1	68.7	+1.5
layer	1.806	44.6	80.3	81.6	+1.4
boil	2.443	42.9	68.7	70.1	+1.4
collective	2.347	39.5	64.3	65.4	+1.1
civilian	1.504	48.7	88.2	88.4	+0.2
provincial	0.293	95.8	96.5	96.5	0.0
overlook	1.597	41.6	86.1	86.1	0.0
impressive	0.000	100.	100	100.	0.0
bucket	1.974	56.8	71.4	71.4	0.0
complain	0.701	87.5	89.7	89.6	-0.1
spite	0.404	94.3	96.5	96.4	-0.2
lemon	2.398	36.3	71.2	70.6	-0.6
literary	1.661	48.7	66.5	65.7	-0.9
connect	2.283	52.7	56.8	55.8	-0.9
attribute	1.949	46.9	76.0	75.0	-1.0
confine	1.392	74.1	83.9	82.8	-1.1
comic	2.033	52.9	74.9	73.8	-1.1
cell	2.099	49.2	74.6	73.5	-1.1
cook	2.386	46.3	77.7	76.4	-1.3
intensify	1.316	51.7	72.8	71.2	-1.5
expression	2.137	36.4	64.0	61.1	-2.9
average	1.874	52.5	75.0	76.4	+1.4

Figure 4: Comparison to Naive Bayes using SENSEVAL dry-run data.

Selected model better	
word	p-value
sick	0.004
curious	0.006
beam	0.016
bake	0.027
brick	0.028
drain	0.048
raider	0.059
impress	0.080
drift	0.084
sugar	0.088

Naive Bayes better	
word	p-value
connect	0.346
comic	0.338
spite	0.334
expression	0.325
literary	0.311
cell	0.283
attribute	0.231
confine	0.172
cook	0.146
intensify	0.107

Figure 5: Statistical significance of gains versus Naive Bayes.

	Response												
Key	_336	_337	_338	_355	_410	_412	_537	_584	_585	_517	_551	_708	
504336	78	0	1	1	3	0	0	1	0	0	0	0	84
504337	1	41	1	3	1	0	0	0	0	3	0	0	50
504338	1	1	37	6	7	1	1	1	0	5	2	0	62
504355	0	3	8	16	0	0	0	0	0	0	0	0	27
504410	0	2	9	1	4	3	0	0	0	0	5	0	24
504412	0	0	5	0	3	4	1	0	0	1	1	0	15
504537	0	0	0	0	1	0	1	1	1	2	0	1	7
504584	0	0	2	0	0	0	0	0	1	0	0	0	3
504585	0	0	0	0	0	0	0	0	13	0	0	0	13
516517	0	1	3	2	0	0	0	0	0	0	0	0	6
516551	0	0	3	0	4	0	0	0	0	0	7	0	14
516708	0	0	2	0	0	0	0	1	0	0	0	1	4
	80	48	71	29	23	8	3	4	15	11	15	2	309

Table 5: Confusion matrix for *shake-p* (excluding infrequent senses).

Franz, A. 1996. *Automatic Ambiguity Resolution in Natural Language Processing: An Empirical Approach*, Berlin: Springer-Verlag.

Friedman N., D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29:131–163.

Kilgarriff, A. 1998. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proc. First International Conference on Language Resources and Evaluation*, pp. 581–588, Granada, Spain, May.

Leacock C., G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proc. ARPA Workshop on Human Language Technology*, Princeton, New Jersey.

Mooney, R. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pp. 82–91.

Ng, H. T., and Lee, H. B. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: an Exemplar-Based Approach. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 40-47.

Ng, H. T. 1997. Getting serious about word sense disambiguation. In *Proc. ANLP-97 Workshop, Tagging Text with Lexical Semantics: Why, What, and How?*, Association for Computational Linguistics SIGLEX, Washington, D.C., April 1997

Pedersen, T. and R. Bruce. 1997. A new supervised learning algorithm for word sense disambiguation. In *Proc. of the 14th National Conference on Artificial Intelligence (AAAI-97)*, Providence, Rhode Island.

Pedersen, T. and R. Bruce. 1998. Knowledge-lean word-sense disambiguation. In *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin.

Wiebe, J., R. Bruce, and L. Duan, 1997. Probabilistic Event Categorization. In *Recent Advances in Natural Language Processing (RANLP-97)*, Tsigov Chark, Bulgaria, Sept. 1997.

Wiebe, J., K. McKeever, and R. Bruce, 1998. Mapping collocational properties into machine learning features. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*. Association for Computational Linguistics SIGDAT, Montreal, Quebec, Canada, August 1998.

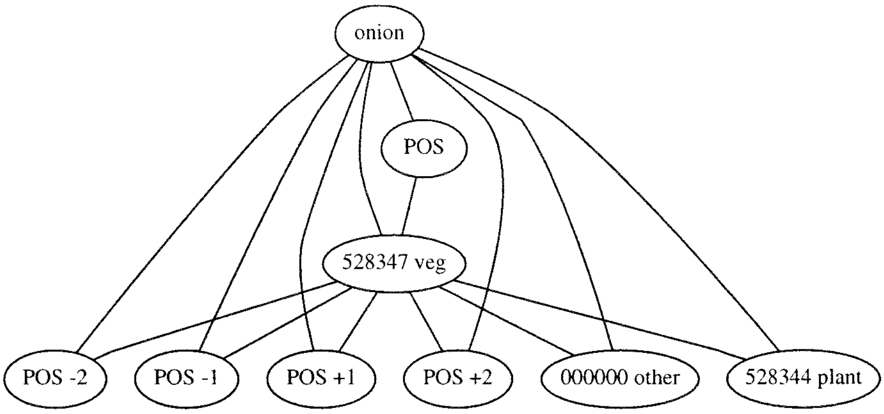


Figure 6: Model selected by backward search for *onion-n*.

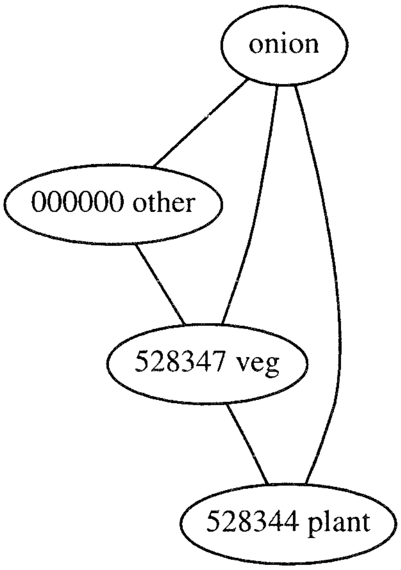


Figure 7: Model selected by forward search for *onion-n*.

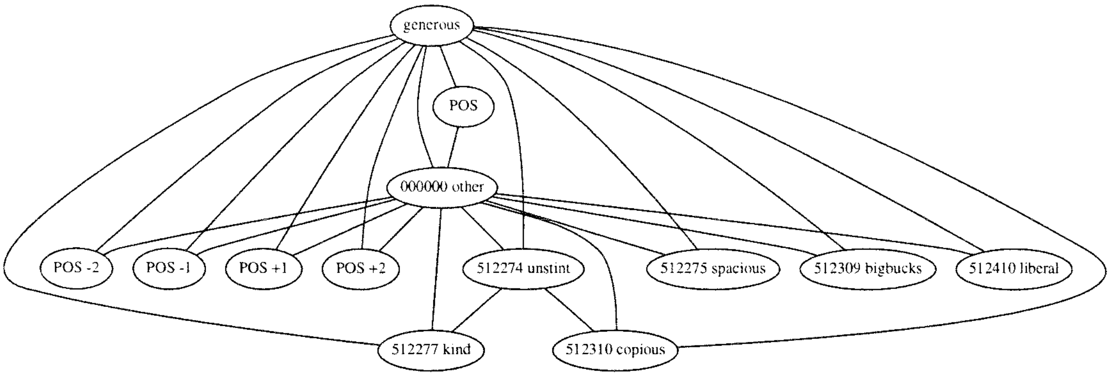


Figure 8: Model selected by backward search for *generous-a*.

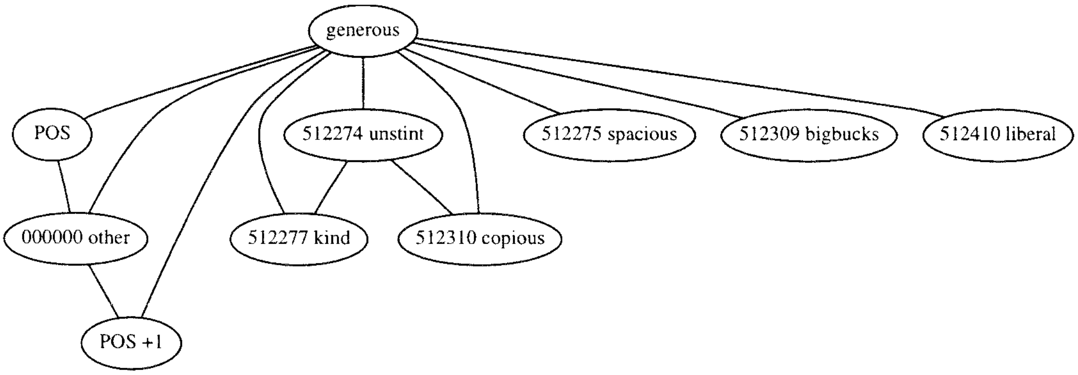


Figure 9: Model selected by forward search for *generous-a*.

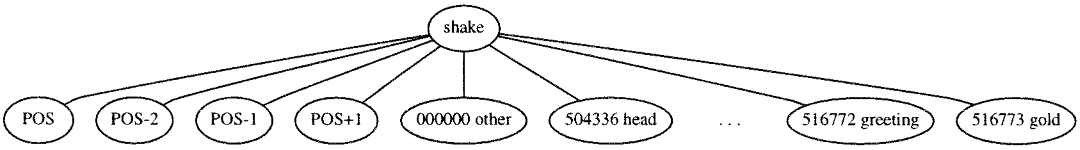
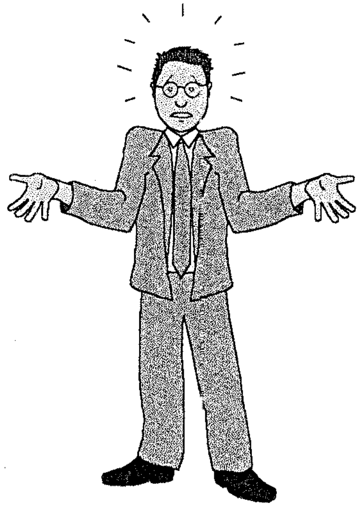


Figure 10: Fixed Naive Bayes model for *shake-p*.

Large Scale WSD applied to Senseval



Durham University

Paul Hawkins - p.m.hawkins@durham.ac.uk

David Nettleton - d.j.nettleton@durham.ac.uk

Introduction

- Motivation was to develop a WSD module to be used in the LOLITA core system.
- LOLITA contains 100,000+ node semantic network which is compatible with WordNet
- Requirements of WSD module are:
  - Large scale - disambiguate all senses of all words.
  - Domain Independent.
  - Disambiguate many ambiguous words in the same sentence.

Further Information

LOLITA is a large scale Natural Language Processing System which has been being developed at Durham University for the last 12 years. It consists of a pipeline architecture in which the main components are Morphology, Parsing, Semantics and Text Generation. This project is aimed at developing a dedicated disambiguation mechanism, which will fit in as part of LOLITA's Semantics. Uncertainty is then carried through the system, however Parsing may eliminate some senses which do not fit a syntactic structure.

Currently the disambiguation module is completely separate from LOLITA so it is not affected by changes in other parts of the system. Identifying Proper Nouns is not considered part of the disambiguation system's role as this has already been developed in LOLITA for MUC. Therefore a Proper Noun sense was not considered for any of the words in Senseval. Also for words where the POS is not given, the disambiguation system has to consider all senses as there is currently no POS tagger to eliminate senses with the wrong POS.

One of the key features of LOLITA is that it is Large Scale and the core analysis can be applied to many NLP tasks. The disambiguation system maintains this feature, and apart from one minor error all sentences in the Senseval evaluation were attempted.

LOLITA's knowledge representation contains the WordNet hierarchy and so the disambiguation algorithm currently uses WordNet senses. By doing so it is able to take advantage of SemCor for training and testing. One aim in developing the disambiguation algorithm is that it can be applied to other lexicons. As the algorithm uses learning then a corpus of training data should be available. If no corpus is available then the system trained on SemCor can be applied if mappings between the lexicon and WordNet exist. This would make the work of use on a wider scale. Senseval was the first opportunity to test both of these features.

Knowledge Sources

Morphology

- Uses frequency information based on the actual word rather than the root form.

Clue Words

- Manually identifies words in the context which will serve as a useful clue.
- The position a clue must appear relative to the ambiguous word can be specified

Further Information

MORPHOLOGY

Actual word frequency information is more specific to the individual problem than frequency information taken from the root word. However in some instances using actual word frequencies can lead to insufficient training data to generate accurate statistics.

For this system using frequency information based on the actual word was particularly useful for words where the POS is not given. For example when trying to disambiguate *shaking* all noun senses of the word will be assigned a zero frequency. The actual word frequency information can go beyond being used only for a primitive POS tagger. The most common sense for *sack* and *sacks* refers to a strong bag, but this sense did not occur in instances when the word is *sacking*. In the evaluation choosing the most common root word for sack gives 50% accuracy, but using the Morphology information increases accuracy to 86.6%. NB This system only achieved 78% for sack due to an error!

CLUE WORDS

Clue words were an add-on to the core system specifically for the Senseval task. To use clue words requires a human to identify useful words in the context and is therefore the one knowledge source which may not be feasible for disambiguating on a large scale lexicon. Despite this, for some words clues are a very valuable knowledge source and they take very little time to find. The position of the clue relative to the ambiguous word can also be specified e.g. knowing whether *hands* appears before or after *shake* helps disambiguation, e.g. *"we shook hands"* and *"his hands were shaking"*. Ideally syntax should be used instead of word position to prevent *"you could sense fear from his shaking hands"* from being disambiguated incorrectly.

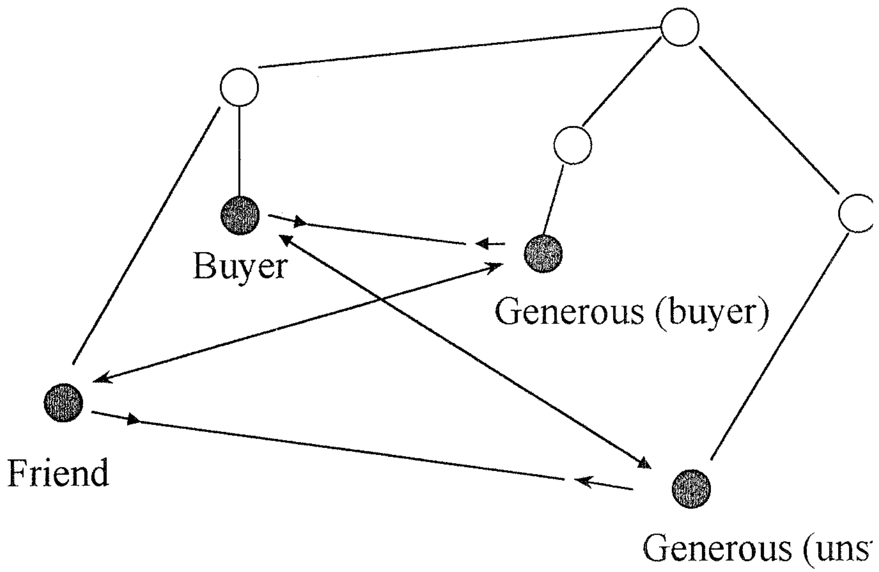
It appears that it is more beneficial to just calculate frequency information for sentences where clue words can not help. This would have benefited disambiguation of *excess* where the most frequent word has a very good clue i.e. "… in excess of …".

Knowledge Sources - Learning

- 2000 nodes are automatically identified in WordNet and scores are stored between each of these nodes.
- The same score is used for all hyponyms of a node.
- This enables words with different POS to be used as context.
- During training scores are adjusted based on the result of disambiguation.

Further Information

Contextual scores between different nodes in WordNet are learnt during training and stored in a matrix. This enables nodes of different POS to be used as context despite there being no path which connects them. Also WordNet was not designed specifically for WSD so just because 2 nodes are close to each other in the WordNet hierarchy doesn't necessarily mean they are useful for disambiguation. For example in *"The buyer made a* ***generous*** *offer"* **generous** is likely to be referring to a different sense to that in *"my friend made a* ***generous*** *offer"*. During training each sentence in the training data is disambiguated, if the disambiguation is incorrect the scores between the context, correct sense and chosen sense are modified. The amount scores are changed is determined by an error function. Increasing a score is represented on the diagram by moving the nodes closer together.



Results

The results are calculated using the fine grained, not minimal algorithm.

	All words	Onion	Generous	Shake
Root Form	57.3	84.6	39.6	23.9
Actual word	61.6	85	37	30.6
Clues	73.7	92.5	44.9	71.1
Training	69.8	85	50.1	61.8
Overall	77.1	92.5	50.7	69.9
Coarse	81.4	92.5	50.7	72.5

Further Information

RESULTS

The system was tailored for the more difficult fine grained evaluation metric. Detailed results are quoted for fine grained.

Using only the actual word frequency and clue words the system obtained an overall accuracy of 73.7% . The results show that adding training information to this increased accuracy by 3.4% over the entire evaluation set. However the training information was only used for 16 words, and for those words training made a 5.2% improvement. The best training mechanism was to initially train on SemCor and the use the Hector data to train further. For *shake,* training information was used, but it proved to reduce accuracy, this is because shake has very good clue words so this knowledge source proved more useful.

In the training data there were only 26 sentences for *onion*, and therefore disambiguation relied purely on clue words.

FUTURE WORK

Other research at Durham is developing a semi-automatic mechanism for adding dictionary definitions to LOLITA's semantic network. The semantics associated with these definitions will provide the information to be able to add selectional restrictions as a knowledge source. Selectional restrictions can suffer from not being able to disambiguate the noun until you know the verb and vice-versa. This problem can be addressed by using selectional restrictions in a combination with the current disambiguation module.

The results of the system have shown the value of heuristics in the form of clue words. An important area of further work is being able to semi-automatically identify clue words thus allowing their application on a larger scale.

The system shows the benefit of combining a learning based approach with a rule based mechanism.

Future Work

- Use information from LOLITA to help disambiguation.
  - Proper Nouns.
  - Identify Subject, Verb and Object to weight importance of context.
  - Use rich semantics for Selectional Restrictions.
- Develop a semi-automatic way of finding clue words.
- Become less dependant on frequency information.

Word Sense Disambiguation Based on The Classification Information Model

Ho Lee and Hae-Chang Rim

Natural Language Processing Lab.

Dept. of Computer Science and Engineering

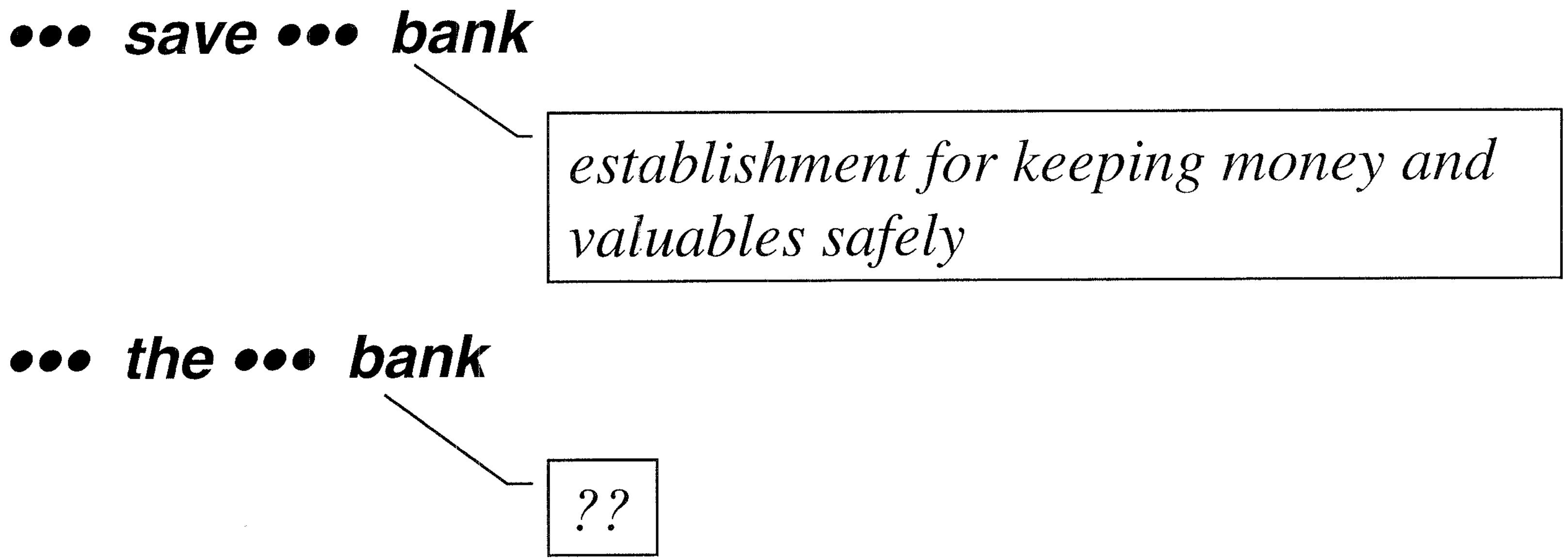
Korea University

Motivation

❑ Hypothesis

☞ the lower entropy an evidence has, the more informative the evidence is

He saves half of his salary every month in the <u>bank</u>.



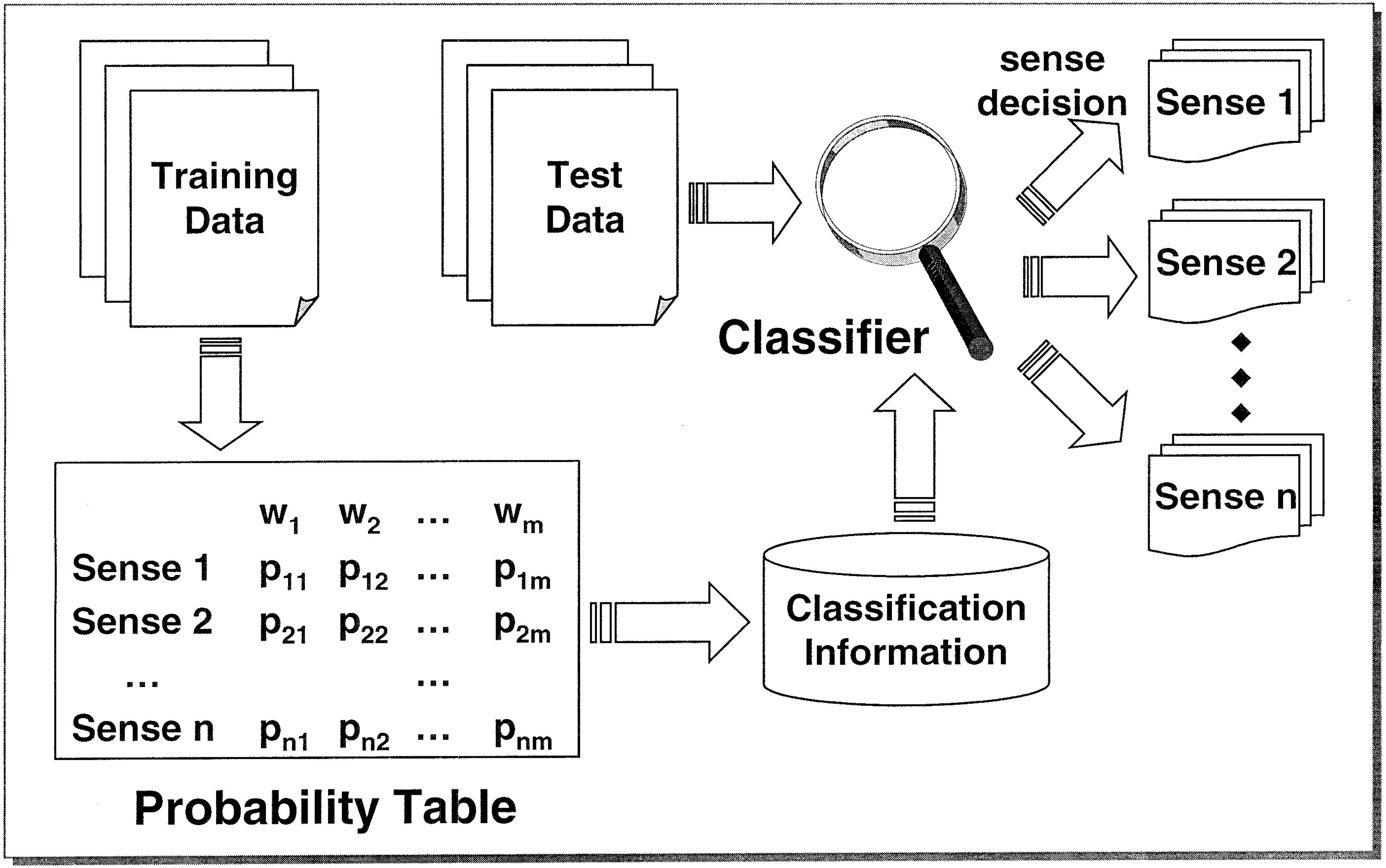
••• save ••• bank

establishment for keeping money and valuables safely

••• the ••• bank

??

System Overview



	w_1	w_2	…	w_m
Sense 1	p_{11}	p_{12}	…	p_{1m}
Sense 2	p_{21}	p_{22}	…	p_{2m}
…			…	
Sense n	p_{n1}	p_{n2}	…	p_{nm}

Probability Table

Classification Information

❑ Components

☞ Most Probable Class(MPC)

- the sense of the target word most closely related to the given evidence

$$MPC_{evidence_k} = \arg\max_{sense_i} p(sense_i \mid evidence_k)$$

☞ Discrimination Score(DS)

- the ability of discriminating senses of the target word

number of sense

$$DS_{evidence_k} = \log_2 n - \left\{ -\sum_{i=1}^{n} p(sense_i \mid evidence_k) \log_2 p(sense_i \mid evidence_k) \right\}$$

maximum entropy

entropy of evidence$_k$

Sense Decision

input sentence : $S = \{evidence_1, evidence_2, \cdots, evidence_m\}$

proper sense in the input sentence

$$MPC(S) = \arg\max_i \sum_{k=1}^{m} DS_{evidence_k}(i)$$

where

$$DS_{evidence_k}(i) = \begin{cases} DS_{evidence_k} & if \quad sense_i = MPC_{evidence_k} \\ 0 & otherwise \end{cases}$$

Example of Sense Decision

He saves half of his salary every month in the bank.

Word	MPC	DS	Sense 1	Sense 2	Sense 3
he	Sense 3	0.1769	0	0	0.1769
save	Sense 1	0.8023	0.8023	0	0
half	Sense 1	0.3299	0.3299	0	0
of	Sense 2	0.1160	0	0.1160	0
his	Sense 3	0.1204	0	0	0.1204
salary	Sense 1	1.1364	1.1364	0	0
every	Sense 1	0.4258	0.4258	0	0
month	Sense 1	0.6731	0.6731	0	0
in	Sense 2	0.2306	0	0.2306	0
the	Sense 2	0.0523	0	0.0523	0
Sum of DS			**3.3675**	0.3989	0.2973

Evidences

❑ Decision list(Yarowsky, 1992)

- ☞ item_{-1} : item immediately to the left
- ☞ item_{+1} : item immediately to the right
- ☞ $\text{item}_{\pm 1}$: item found in ± k word window
- ☞ (item_{-2} ,item_{-1}) : pair of items at offset -2 and -1
- ☞ (item_{-1} ,item_{+1}) : pair of items at offset -1 and +1
- ☞ (item_{+1} ,item_{+2}) : pair of items at offset +1 and +2

❖ items

- surrounding words
- parts-of-speech

Experimental Results - 1

❑ Overall performance(precision)

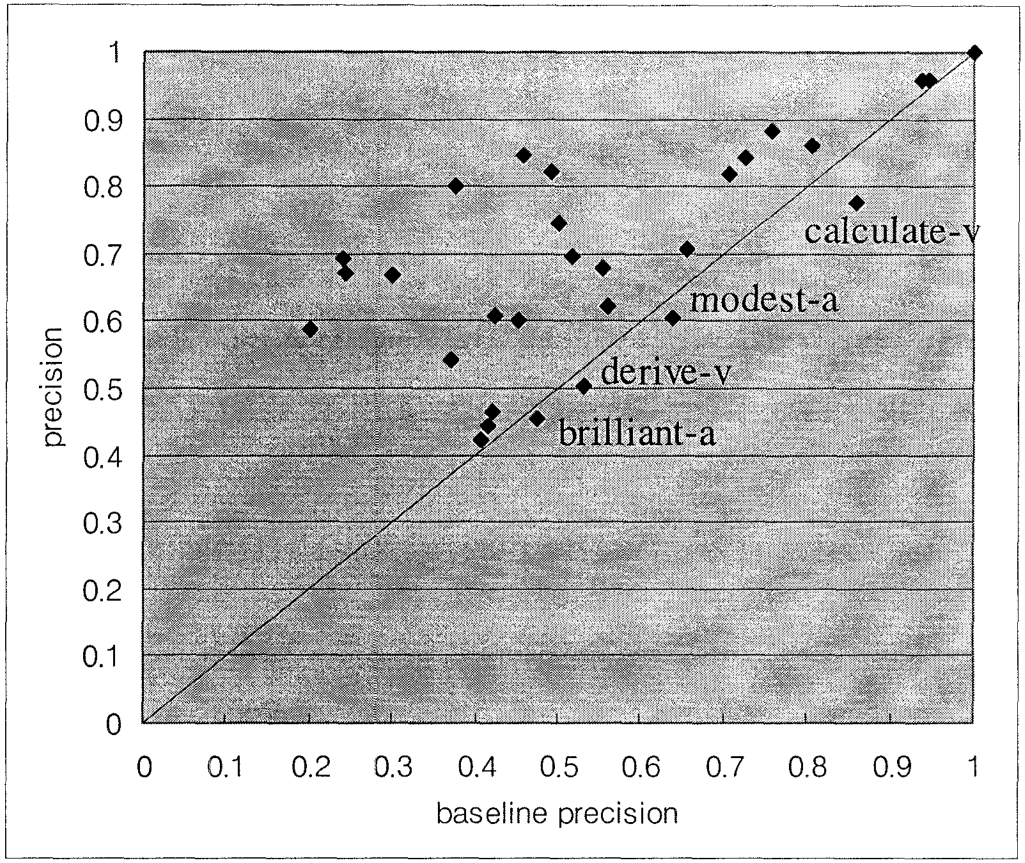
	nouns	*verbs*	*adj.s*	*total*
fine-grained	0.771	0.642	0.674	0.701
mixed-grained	0.825	0.683	0.723	0.740
coarse-grained	0.849	0.695	0.727	0.752

- After fixing sense mapping errors, the results are re-scored e.g. sense 1 of 'bet-v' is mapped to UID 51994('bet-n')
- The system tries to decide senses for all instances of words except the words without training data
  ⇨ recall = precision

Experimental Results - 2

❑ Comparison with baseline method

☞ baseline method : always select the most frequent sense



Summary

❑ Summary

- ☞ a supervised learning model based on the classification information
- ☞ represents evidence by means of decision lists
- ☞ exploits surrounding words and their parts-of-speech

❑ Future works

- ☞ use class-based probability instead of word-based probability
  - overcome data sparseness problem
  - currently applied to Korean
- ☞ combine the classification information with unsupervised learning method
  - prevent knowledge acquisition bottleneck