University of Brighton

ITRI-00-20

Framework and results for English SENSEVAL

Adam Kilgarriff and Joseph Rosenzweig

April, 2000

Also published in Computers and the Humanities 34 (1–2), Special Issue on SENSEVAL, pp 15–48.

Supported by the EPSRC under Grants K18931 and M54971

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: `firstname.lastname@itri.brighton.ac.uk`
FAX: +44 1273 642908 NET: `http://www.itri.brighton.ac.uk`

Framework and Results for English SENSEVAL

A. Kilgarriff
ITRI, University of Brighton

J. Rosenzweig
University of Pennsylvania

Abstract.

SENSEVAL was the first open, community-based evaluation exercise for Word Sense Disambiguation programs. It adopted the quantitative approach to evaluation developed in MUC and other ARPA evaluation exercises. It took place in 1998. In this paper we describe the structure, organisation and results of the SENSEVAL exercise for English. We present and defend various design choices for the exercise, describe the data and gold-standard preparation, consider issues of scoring strategies and baselines, and present the results for the 18 participating systems. The exercise identifies the state-of-the-art for fine-grained word sense disambiguation, where training data is available, as 74–78% correct, with a number of algorithms approaching this level of performance. For systems that did not assume the availability of training data, performance was markedly lower and also more variable. Human inter-tagger agreement was high, with the gold standard taggings being around 95% replicable.

Keywords: word sense disambiguation, evaluation, SENSEVAL

1. Introduction

In this paper we describe the structure, organisation and results of the SENSEVAL exercise for English.

The architecture of the evaluation was as in MUC and other ARPA evaluations (Hirschman, 1998). First, all likely participants were invited to express their interest and participate in the exercise design. A timetable was worked out. A plan for selecting evaluation materials was agreed. Human annotators were set on the task of generating a set of correct answers, the 'gold standard'. The gold standard materials, without answers, were released to participants, who then had a short time to run their programs over them and return their sets of answers to the organisers. The organisers then scored the answers, and the scores were announced and discussed at a workshop.

Setting up the exercise involved a number of choices — of task, corpus and dictionary, words to be investigated and relation to word class tagging. In sections 2–5, we describe the theoretical and practical considerations and the choices that were made.

In the following sections we describe the data, the manual tagging process (including an analysis of inter-tagger agreement), the scoring regime, the participating systems, and the baselines. Section 11 presents the results. Section 12 considers the relations between polysemy, entropy and task difficulty, and section 13, an experiment in pooling the results of different systems.

The first three Appendices briefly describe the three systems for which there is no full paper in the Special Issue, and the fourth presents samples of the dictionary entries and corpus instances used in SENSEVAL.

A note on terminology: in the following, a 'corpus instance' or 'instance' is an instance of a word occurring in context in a corpus, or, a particular token of the word. A 'word' is a word type, or lexical word. Thus the sentence *Dog eats dog* contains two, not three, words.

2. Choice of task: 'all-words' *vs.* 'lexical-sample'

Evidently, the task was word sense disambiguation (WSD), in English. Two variants of the WSD task are 'all-words' and 'lexical-sample'. In all-words, participating systems have to disambiguate all words (or all open-class words) in a set of texts. In lexical-sample, first, a sample of words is selected. Then, for each sample word, a number of corpus instances are selected. Participating systems then have to disambiguate just the sample-word instances.

For SENSEVAL, the lexical-sample variant was chosen. The reasons were linked with issues of dictionary choice and corpus choice. They included the following:

- Cost-effectiveness of tagging: it is easier and quicker for humans to sense-tag accurately if they concentrate on one word, and tag multiple occurrences of it, than if they have to focus on a new dictionary entry for each word to be tagged.

- The all-words task requires access to a full dictionary. There are very few full dictionaries available (for low or no cost) so dictionary choice would have been severely limited. The lexical-sample task required only as many dictionary entries as there were words in the sample.

- Many of the systems interested in participating could not have participated in the all-words task, either because they needed sense-tagged training data (see also below) or because the needed some manual input to augment the dictionary entry for each word to be disambiguated.

- It would be possible for systems designed for the all-words task to participate in the lexical-sample task, whereas the converse was not possible (except for a hopelessly small subset of the data). A system that tags all words does, by definition, tag a subset of the words.

- Provided the sample was well-chosen, the lexical-sample strategy would be more informative about the current strengths and failings of WSD research than the all-words task. The all-words task would provide too little data about the problems presented by any particular word to sustain much analysis.[1]

2.1. A QUESTION OF TIMING

All-words systems can participate in the lexical-sample task, but at a disadvantage. The disadvantage would be substantially offset if the words in the lexical sample were not announced prior to the distribution of the evaluation material. Then, it would be possible for supervised learning systems to participate and to exploit training materials, but there would not be time for non-automatic tailoring of systems to the particular problems presented by the words in the sample. This strategy was considered, and was partially adopted, with the words being announced just two weeks (in principle) before the test data was released. The constraints on its adoption were both practical and theoretical:

- Systems such as CLRES and UPC-EHU[2] perform extensive analyses of dictionary definitions the software needs to be adapted to work with the particular dictionary format. For these systems to participate, a substantial sample of entries was required for porting the system to the new dictionary. To this end, a set of 'dry run' dictionary entries was distributed early. It was however possible that the forty lexical entries in the dry-run sample did not exhibit the full range of dictionary-formatting phenomena found in the thirty-five evaluation sample entries.

- The organisers did not share the assumption of some researchers that manual input, for the lexical entry of each word to be disambiguated, should be viewed as illegitimate. One high-performing system (DURHAM) owed some of its accuracy to what was, in effect,

[1] For a fuller statement of the case see (Kilgarriff, 1998). For the counter-arguments, see (Wilks, this volume).

[2] All systems are referred to by their short names, as given in Table IV.

additional lexicography undertaken for the words in the evaluation sample. (Harley and Glennon, 1997) describes a high-quality WSD system built on the basis of telling lexicographers to put into the dictionary, the information that would be required for WSD. The objection to this approach is economic: there are vast numbers of ambiguous words, so it is too expensive. That need not be so. As (Moon, this volume) shows, the number of words requiring disambiguation in English is in the order of 10,000: if each requires fifteen minutes of human input, the whole lexicon calls for around two person-years, which is no more than many WSD systems have taken to design and build.

The customer for a WSD system will be interested in its performance, not the purity of its knowledge-acquisition methods.

- In practice, it was not viable to draw a line between legitimate 'debugging' and possibly illegitimate 'manual system enhancement'. Nor was it possible to set the deadlines very tightly, given the usual complications of conflicting deadlines, absences from the office, etc. 'Manual system enhancement' could not be severely constrained by time limits.

3. Choice of dictionary and corpus

The HECTOR lexical database was chosen. HECTOR was a joint Oxford University Press/Digital project (Atkins, 1993) in which a database with linked dictionary and corpus was developed. For a sample of words, dictionary entries were written in tandem with sense-tagging all occurrences of the word in a 17M-word corpus (a pilot for the British National Corpus[3]). The sample of words comprised those items with between 300 and 1000 instances in the corpus. The tagger-lexicographers were highly skilled and experienced. There was some editing, with a second lexicographer going through the work of the first, but no extensive consistency checking.

The primary reason for the choice was a simple one. At the time when a choice was needed, it was not evident whether there was any funding available for manual tagging. Had funding not been forthcoming, then, with the HECTOR data, it would still have been possible to run SENSEVAL as corpus instances had been manually tagged in the HECTOR project. (In the event, there was funding,[4] and all evaluation data was doubly re-tagged. Un-re-tagged HECTOR data was used

[3] Hereafter the BNC: for more information see `http://info.ax.ac.uk/bnc`

[4] The funding was from the UK EPSRC under grant M03481.

for the training dataset.) The resource has been offered for use under licence in SENSEVAL, without charge, by Oxford University Press.

There was one other possible source of already tagged data: the SEMCOR corpus, tagged according to WordNet senses (Fellbaum, 1998). However, SEMCOR was already widely used in the WSD community so it could not provide 'unseen' data for evaluation. Also, it had been tagged according to an all-words strategy, so would have pointed to an all-words evaluation.

Supplementary reasons for choosing the HECTOR data were:

- The dictionary entries were fuller than in most paper dictionaries or WordNet, and this was likely to be beneficial for WSD.

- The lexicography was highly corpus-driven, and was thus (arguably) representative of the kind of lexicography that is likely to serve NLP well in the future.

- No previous WSD work had used HECTOR, so no WSD team was at a particular advantage.

- The corpus was of general English. It had been decided at a previous ACL SIGLEX meeting (Kilgarriff, 1997) that WSD evaluation should aim to use general language rather than a specific domain.

One disadvantage of the HECTOR corpus material in the form in which it was received from OUP was that corpus instances were associated with very little context: generally two sentences and sometimes just one sentence. Strategies for gleaning information from a wider context would not show their strength.

4. Lexicon sampling

A criticism of earlier forays into lexical-sample WSD evaluation is that the lexical sample had been chosen according to the whim of the experimenter (or to coincide with earlier experimenters' selections). For SENSEVAL, a principled approach based on a stratified random sample was used. A simple random sample of polysemous words would have been inappropriate, since, given the Zipfian distribution of word frequencies, most or all of the sample would have been of low-frequency words. High frequency words are both intrinsically more significant (as they account for more word tokens) and tend to present a more challenging WSD problem (as there is a high correlation between frequency and semantic complexity).

For English SENSEVAL, a sampling frame was devised in which words were classified according to their frequency (in the BNC) and their polysemy level (in WordNet). For each word class under consideration (noun, verb, adjective), frequency and polysemy were divided into four bands, giving a 4×4 grid. A sample size of 40 words was then set (for both dry-run and evaluation samples). The sample was divided between the grid cells according to: (1) the number of words in the grid and (2) the proportion of corpus tokens they accounted for. We were constrained to use HECTOR words so we then took a random sample of the required size of the HECTOR words in each grid cell. (For some grid cells, there were not enough HECTOR words, so substitutes were taken from other cells.)[5]

The number of gold-standard corpus instances per word was also based on the grid. For simpler words (with lower frequency and polysemy) a smaller number was appropriate. Higher-frequency or more polysemous words tend to be more complex and harder for WSD so more data was needed. Different grid-cells were associated with different numbers of corpus-instances-per-word-type, from 160, for the least common and polysemous words, to 400, for the most.

5. Gold-standard specifications

5.1. Word Class (and Part-of-Speech Tagging): Words and Tasks

Word class issues complicated the task definition. The primary issue was: was the assignment of word class (POS-tagging) to be seen as part of the WSD task? In brief, the argument **for** was that, in any real application, the word sense tagging and POS-tagging will be closely related, with each potentially providing constraints on the other. The argument **against** was 'divide and rule': POS-tagging is a distinct subarea of NLP, with its own strategies and issues, and (arguably) a high accuracy rate, so was best kept out of the equation, the better to focus on WSD performance. A previous SIGLEX meeting had seen a majority in favour of decoupling, but no unanimity.

For English SENSEVAL, for most of the evaluation words, the tasks were decoupled, with the part-of-speech (noun, verb or adjective) of the corpus instance specified by the organisers as part of the input to the WSD task. However for five words, the tasks were not decoupled, so participating systems had to assign a sense without prior knowledge of word class. This gave rise to a distinction between words and 'tasks'.

[5] The sampling strategy is fully described in (Kilgarriff, 1998).

Each SENSEVAL **task** was identified by a word and either a word class (noun, verb or adjective) or **p** for 'Part of speech not provided'. The task name comprised the word and one of **-n**, **-v**, **-a** or **-p**.

Some words were associated with more than one task, eg. **sack** has **sack-n** and **sack-v**.[6] Thus there are both words that occur with different parts of speech in different tasks, and words that occur with unspecified part of speech in a single **-p** task. The evaluation sample comprised 34 words and 41 tasks.[7]

The manual taggers assigned word class as well as sense tag so that, for example, a corpus instance of *sack* could be allocated to either the **sack-n** or **sack-v** task. Most of the time this was straightforward but there were exceptions, notably gerunds (*his* **sanctioning** *of the initiative*), participles (*severely* **shaken** *he* ...) and modifiers (**bitter** *beer*).

Gerund instances were taken out of the **-v** tasks, as they were not verbal. Participles and nominal modifiers revealed a deeper issue. It was a useful simplifying assumption that lexical word class matched corpus-instance word class, but there were exceptions. Thus verbal *float* had a 'sound' sense, "to be heard from a distance", and adjectival *floating* had no corresponding sense, yet the instance

> the *floating* melody reached even the Vizier's ears

was clearly an adjectival use of the 'sound' sense. In the gold standard there are a very small number of instances where there is a mismatch between the word class of the corpus instance, and the word class of the semantically closest word sense.

5.2. Proper Names

Straightforward proper-name instances were not included in the gold standard materials. There were however also a number of instances where the word was being used in one of its standard senses **within** a proper name. Thus the *Cheltenham Hurdle* is a hurdle race, and *Brer Rabbit* is a rabbit. These cases were included in the gold standard, with the complete correct answer having two parts: the appropriate sense for **hurdle** or **rabbit**, and the proper-name tag, PROPER, which was available for all words.

[6] This was motivated by economy: it made an extra pass over the data to determine part-of-speech unnecessary.

[7] **float** was associated with three tasks, **float-v**, **float-n** and **floating-a**, sometimes also called **float-a**.

5.3. Other difficult cases

For cases where more than one word sense applied, or appeared equally valid, or there was insufficient context to say which applied, the gold standard specifies all salient senses. Where none of the HECTOR senses fit, the gold standard states "unassignable" with the universally-available tag UNASS. For 'exploitations', where the use is related to one of the senses in some way but does not directly match it, the gold standard specifies both the sense and UNASS. (In the taggers' first pass, there was a finer-grained analysis of the misfit categories, but for WSD evaluation, a scheme simple enough to score by was required.) For the taggers' perspective on the exercise, and the instances that made the work difficult and interesting, see (Krishnamurthy and Nicholls, this volume).

6. The data

There were three data distributions. The target dates were

end April Dry run data

end June Training data

mid July Evaluation data

6.1. Dry-run data

The dry-run data comprised lexical entries and hand-tagged corpus instances, and was sampled in the same way as the training and evaluation data. It could be used to adapt systems to the format and style of data that would be used for evaluation. It comprised the words and associated numbers of instances shown in Table 6.1.

6.2. Training data

The training distribution comprised lexical entries and hand-tagged corpus instances for the lexical sample that was to be used for evaluation. The lexical entries were provided so that participants could ensure that their systems could parse and exploit the dictionary entries and add to them where necessary (see discussion on timing above). The corpus instances were provided so that supervised-training systems could be trained for the words in the lexical sample. For five words there was no training data (see Table II), and for the remainder, the quantity

Table I. Dry run data: words and numbers of instances

attribute	364	bake	346	beam	337	boil	567
brick	586	bucket	174	cell	698	civilian	582
collective	495	comic	502	complain	1116	confine	586
connect	516	cook	1922	creamy	101	curious	465
dawn	551	drain	578	drift	515	expression	917
govern	593	impress	641	impressive	711	intensify	234
layer	492	lemon	225	literary	690	overlook	437
port	874	provincial	373	raider	164	sick	639
spite	577	storm	763	sugar	855	threaten	307
underground	519	vegetable	636				

varied widely between 26 and 2008 instances, depending simply on how many there were available.

In both dry-run and training data, corpus instances were provided complete with the sense-tag that had been assigned as part of the original HECTOR tagging, but there had been no re-tagging. Unlike the evaluation data, there was no explicit information on word class, though this was deducible from the sense-tag with over 99% accuracy.[8]

6.3. Evaluation data

The evaluation distribution simply contained a set of corpus instances for each task. Each had been tagged by at least three humans, though these tags were, of course, not part of the distribution. (It did not contain lexical entries because they were already available in the training distribution.)

Examples of lexical entries and corpus instances are included in Appendix 4. Lexical entries were distributed in their native format, minimally-structured SGML, with a utility to convert into latex and thereby to produce output of the form shown in Appendix 4. Corpus entries were distributed as ASCII texts, with the word to be tagged indicated by a <tag> tag, each instance having a six-digit reference number (starting with 7, unique within a given task), one sentence on each line, and instances separated by an empty line.

There were 8448 corpus instances in total in the evaluation data. The tasks and associated quantities of data are presented in Table II.

[8] In the event, there were some differences of format between the dry-run training data, and evaluation data, because, between the releases, there was more time to

Table II. Evaluation tasks and dataset sizes [1] Multiple tasks for these words: training data shared [2] No training data for these items

Nouns		Verbs		Adjectives		Indeterminates	
-n		**-v**		**-a**		**-p**	
accident	267	amaze	70	brilliant	229	band	302
behaviour	279	bet[1]	177	deaf[2]	122	bitter	373
bet[1]	274	bother	209	floating[1]	47	hurdle[2]	323
disability[2]	160	bury	201	generous	227	sanction	431
excess	186	calculate	217	giant[1]	97	shake	356
float[1]	75	consume	186	modest	270		
giant[1]	118	derive	216	slight	218		
knee	251	float[1]	229	wooden	195		
onion	214	invade	207				
promise[1]	113	promise[1]	224				
rabbit[2]	221	sack[1]	178				
sack[1]	82	scrap[1]	186				
scrap[1]	156	seize	259				
shirt	184						
steering[2]	176						
TOTAL	2756	TOTAL	2501	TOTAL	1406	TOTAL	1785

6.4. WordNet mapping

For participants whose systems output WordNet senses, a mapping from WordNet senses to HECTOR senses was provided. As previous evidence of sense-mapping has always found (eg. (Byrd et al., 1987)) the result is not altogether satisfactory, with gaps, one-to-many and many-to-many mappings.

6.5. Specifications for Returning Results

Systems were required to return, for scoring, a one-line answer for each corpus instance for which they were returning a result. A line comprised

1. The task
2. The reference number for the instance

clean up data and to work on the task specification. This caused some participants substantial inconvenience.

3. One or more sense tags, optionally with associated probabilities. Where there were no numbers, the probability mass was shared between all listed tags.

7. Gold Standard Preparation: manual tagging

The preparation of the gold standard included:

- obtaining funding to pay taggers
- selecting individuals
- selection of materials, including weeding-out anomalous items[9]
- preparation of detailed tagging instructions, including fine-grained definition of the evaluation task in relation to eg., word class, proper names, hard-to-tag cases, and data formats for distributing work to taggers and for them to return their answer keys
- sending out data to taggers
- processing returned work to identify those cases where there was unanimity amongst taggers, and those where there was not (so arbitration was required)
- administration of arbitration phase.

All stages were completed between March and August 1998.

7.1. Inter-tagger agreement and replicability

Preparation of a gold standard worthy of the name was critical to the validity of the whole SENSEVAL exercise. The issue is discussed in detail in (Gale et al., 1992) and (Kilgarriff, 1998). A gold standard corpus must be replicable to a high degree: the taggings must be correct, and it can only be deemed that they are correct if different individuals or teams tagging the same instance dependably arrive at the same tag. Gale et al. identify the problem as one of identifying the 'upper bound' for the performance of a WSD program. If people can only agree on the correct answer $x\%$ of the time, a claim that a program achieves more

[9] For example, numerous corpus instances had been used as HECTOR dictionary examples. These needed weeding out from the evaluation materials. With thanks to Frédérique Segond and Christiane Fellbaum for pointing this out.

than $x\%$ accuracy is hard to interpret, and $x\%$ is the upper bound for what the program can (meaningfully) achieve.

There have been some discussions as to what this upper bound might be. Gale et al. review a psycholinguistic study (Jorgensen, 1990) in which the level of agreement averaged 68%. But an **upper** bound of 68% is disastrous for the enterprise, since it implies that the best a program could possibly do is still not remotely good enough for any practical purpose. Even worse news comes from (Ng and Lee, 1996), who re-tagged parts of the manually tagged SEMCOR corpus (Fellbaum, 1998). The taggings matched only 57% of the time. For SENSEVAL, it was critical to achieve a higher replicability figure. To this end, the individuals to do the tagging were carefully chosen: whereas other tagging exercises had mostly used students, SENSEVAL used professional lexicographers. A dictionary which would facilitate accurate tagging was selected. Taggers were encouraged to give multiple tags (one of which might be UNASS) rather than make a hard choice, where more than one tag was a good candidate. And the material was multiply tagged, and an arbitration phase introduced. First, two or three lexicographers provided taggings. Then, any instances where these taggings were not identical were forwarded to a further lexicographer for arbitration.

At the time of the SENSEVAL workshop, the tagging procedure (including arbitration) had been undertaken once for each corpus instance. Individual lexicographers' initial pre-arbitration results were scored against the post-arbitration results. The scoring algorithm was as for system scores. The scores ranged between 88% to 100%, with just five out of 122 results for <lexicographer, word> pairs falling below 95%.

To determine the replicability of the whole process in a thoroughgoing way, the exercise was repeated for a sample of four of the words. The words were selected to reflect the spread of difficulty: we took the word which had given rise to the lowest inter-tagger agreement in the previous round, (*generous*, 6 senses), the word that had given rise to the highest, (*sack*, 12 senses), and two words from the middle of the range (*onion*, 5, and *shake*, 36). The 1057 corpus instances for the four words were tagged by two lexicographers who had not seen the data before; the non-identical taggings were forwarded to a third for arbitration. These taggings were then compared with the ones produced previously.

Table 7.1 shows, for each word, the number of corpus instances (Inst), the number of multiply-tagged instances in each of the two sets of taggings (A and B), and the level of agreement between the two sets (Agr).

Table III. Replicability of manual tagging

Word	Inst	A	B	Agr %
generous	227	76	68	88.7
onion	214	10	11	98.9
sack	260	0	3	99.4
shake	356	35	49	95.1
ALL	1057	121	131	95.5

There were 240 partial mismatches, with partial credit assigned, in contrast to just 7 complete mismatches. For evidence of the kinds of cases on which there were differences of taggings, see Krishnamurthy and Nicholls (this volume).

This was a most encouraging result, which showed that it was possible to organise manual tagging in a way that gave rise to high replicability, thereby validating the WSD enterprise in general and SENSEVAL in particular.

8. Scoring

Three granularity levels for scoring were defined. At the fine-grained level, only identical sense tags counted as a match. At the coarse-grained level, all subsense tags (corresponding to codes such as 1.1, 2.1) were assimilated to main sense tags (corresponding to codes such as 1, 2) in both the answer file and in the key file, so a guess of 1.1 in the answer file counts as an exact match of a correct answer of 1, 1.1 or 1.2 in the key. At the third, 'mixed-grain' level, full credit for a guess was awarded if it was subsumed by an answer in the key file, and partial credit if it subsumed such an answer, as described in Melamed and Resnik (this volume; hereafter MR).

For many instances in HECTOR, it does seem appropriate to give credit for a sense when the correct answer is a subsense of that sense, and *vice versa* —but in others it does not. Consider HECTOR's sense 1 of *shake*, MOVE, defined as:

> to move (someone or something) forcefully or quickly up and down

Sense 1.1 CLEAN, is,

> to remove (a substance, dirt, object etc.) from something by agitating it

and it does seem appropriate to give credit where sense 1.2 is given for 1 or *vice versa*. But sense 1.2, DUST, is

> to leave that place or abandon that thing for ever

as in *shaking the dust of Kingston off her feet forever*. While the etymological link to senses 1 and 1.1 is evident, the difference in meaning is such that it seems quite inappropriate to assign credit to a guess of 1.2 where the correct answer was 1. The validity of subsuming subsenses under main senses remains open to question.

In the event, the choice of scoring scheme made little difference to the relative scores of different systems, or of systems on different tasks. Except where explicitly noted, the remainder of the paper refers only to fine-grained scores.

Where a system returned several answers, it was assumed that the probability mass was shared between them, and credit was assigned as described in MR.[10] All the scoring policies make the MR assumption that there is exactly one correct answer for each instance. This is so even though provision is made for multiple answers in the answer key, because these answers are viewed disjunctively, that is, the interpretation is that any of them could be the correct answer, not that the correct answer comprises all of them. It is hard to determine on a general basis whether a given instance of multiple tags in the key should be interpreted conjunctively or disjunctively (see also Calzolari and Corazzari, this volume).

The precision or performance of a system is computed by summing the scores over all test items that the system guesses on, and dividing by the number of guessed-on items. Recall is computed by summing the system's scores over all items (counting unguessed-on items as a zero score), and dividing by the total number of items in the evaluation dataset or subtask of evaluation. These measures may be viewed as the expected precision and recall of the system in a simpler testing situation where only one answer for each question may be returned, and where each answer either matches the key exactly or does not match it at all.[11]

[10] If the numbers associated with multiple guesses that a system returned did not sum to one, they were first normalised so that they did.

[11] There was one further variable in the scoring: 'minimal' vs 'full' scoring. Minimal scoring was defined as the score a system achieved if it was evaluated only on those instances where the key was a single sense. The intention was to provide a score with a clear, unequivocal interpretation. In the event, once again, the choice of scheme made little difference to the relative scores and the remainder of the paper refers only to full scores.

9. Systems

The 18 systems which returned results are shown in Table IV.[12]

Systems differ greatly in terms of the input data they require and the methodology they employ. This makes comparisons particularly odious, but, to make the comparisons marginally more palatable, they were classified into two broad categories, the supervised systems, which needed sense-tagged training instances of each word they were to disambiguate, and the ones which did not, hereafter 'unsupervised'.[13]

The scheme is a first pass, and various classifications seem anomalous. Some supervised systems are also equipped to fall back on alternative tagging strategies in the absence of an annotated training corpus, while some unsupervised systems default to a frequency-based guess if information from a training corpus is available. Systems such as SUSS and CLRES were in principle unsupervised, but used the training data (as well as the dry-run data) to debug and improve the configuration of their programs. We use the scheme to simplify the presentation of results, but ask the reader to treat it indulgently.

All systems are described by their authors in this Special Issue, either in a paper, or, for CUP-CLS, MALAYSIA and OTTAWA, in Appendices to this paper.

9.1. UPPER BOUND USING WORDNET MAPPING

Four of the systems (UPC-EHU-UN, UPC-EHU-SU, SUSSEX AND OTTAWA) disambiguated according to WordNet senses and used the HECTOR–WordNet map provided by the organisers. To assess how system performance was degraded by the mapping, we computed an upper bound by taking the gold-standard answers, mapping them to the WordNet tags (using an inverted version of the same mapping) and then mapping them back to HECTOR tags. The resulting tagging was scored using the standard scoring software. The strategy gave answers for just 79% of instances; for the remaining 21%, the correct HECTOR tag did not feature in the mapping. Precision was also 79%. Even though the set of tags is guaranteed to include all correct tags, on this algorithm, the mappings in both directions are frequently one-to-many so the correct

[12] All but one returned results before the workshop. Several returned further results by a later, post-workshop deadline. CUP-CLS was the one system that only returned results by the later date.

[13] Earlier classifications made a further distinction within the unsupervised systems, between the 'all-words' systems that could disambiguate all (content) words, and 'others', which could not. In the event this distinction was hard to draw, and there was only one likely candidate for this 'other' category, so the distinction is not used here.

Table IV. Participating systems for English

Group	Contact	Shortname
Unsupervised		
CL Research, USA	Litkowski	clres
Tech U Catalonia, Basque U	Agirre	upc-ehu-un
U Ottawa	Barker	ottawa
U Manitoba	Lin	mani-dl-dict
U Sunderland	Ellman	suss
U Sussex	McCarthy	sussex
U Sains Malaysia	Guo	malaysia
XEROX-Grenoble, CELI, Torino	Segond	xeroxceli
Post-workshop results only		
CUP/Cambridge Lang Services	Harley	cup-cls
Supervised		
Bertin, U Avignon	de Loupy	avignon
Educ Testing Service, Princeton	Leacock	ets-pu
John Hopkins U	Yarowsky	hopkins
Korea U	Ho Lee	korea
New Mex State, UNC Asheville	O'Hara	grling-sdm
Tech U Catalonia, Basque U	Agirre	upc-ehu-su
U Durham	Hawkins	durham
U Manitoba	Suderman	manitoba-ks
U Manitoba	Lin	manitoba-dl
U Tilburg	Daelemans	tilburg

answer is diluted. Evidently, systems using the WordNet mapping were operating under a severe handicap and their performance cannot usefully be compared with that of systems using HECTOR tags directly.[14] (Other systems such as ETS-PU and the two MANITOBA systems used WordNet or other lexical resources, but not in ways which left them crucially reliant on the sense-mapping.)

[14] CUP-CLS was under a similar handicap, as it used a mapping for the CIDE dictionary.

10. Baselines

Two sets of baselines are used: those that make use of the corpus training data, and those that only make use of the definitions and illustrative examples found in the dictionary entries for the target words. The baselines which use training data are intended for comparison with supervised systems, while the ones that use only the dictionary are suitable for comparisons with unsupervised systems.

None of the baselines in either set draws on any form of linguistic knowledge, except for those that are coupled with the phrase filter, which recognises inflected forms of words and applies rudimentary ordering constraints for multi-word expressions. The baselines, like the systems, are free to exploit the pre-specified part-of-speech tags of the words to be disambiguated for the noun, verb and adjective (hereafter **-nva**) tasks. Some of the baselines also make use of the root forms of the words to be disambiguated.[15]

The baselines used for comparison in this paper are:

RANDOM:
—gives equal weight to all sense tags that match the test word's root form and, for **-nva** tasks, part of speech.[16]

COMMONEST:
—always selects the most frequent of the training-corpus sense tags that match the test word's root form (and, for **-nva** tasks, part of speech). The frequency calculation ignores cases involving multiple sense tags or where the tag is PROPER or UNASS. It makes no guesses on the words for which no training data was available.

LESK:
—uses a simplification of the strategy suggested by (Lesk, 1986), choosing the sense of a test word's root whose dictionary definition and example texts have the most words in common with the words around the instance to be disambiguated. The strategy is, for each word to be tagged:

```
(a) For each sense s of that word,
```

[15] The root form is given as the prefix of the file name that a test item occurs in, so is, in this exercise, available to all systems. If it were not given in the file name, some linguistic analysis would be required to obtain it.

[16] Here and for other comparable computations below, PROPER and UNASS tags are left out, since giving them equal weight would greatly reduce the weights for actual dictionary senses of low-polysemy words.

```
(b)   set weight(s) to zero.
(c) Identify set of unique words W in surrounding sentence.
(d) For each word w in W,
(e)   for each sense s of the word to be tagged,
(f)     if w occurs in the definition or example sentences of s,
(g)       add weight(w) to weight(s).
(h) Choose the sense with greatest weight(s)
```

Weight(w) is defined as the inverse document frequency (IDF) of the word w over the definitions and example sentences in the dictionary. IDF is a standard measure used in information retrieval which serves to discount function words in a principled way, since it is inversely proportional to a word's likelihood of appearing in an arbitrary definition or example. The IDF of words like *the, and, of* is low, as they appear in most definitions, while the IDF of content words is high. The IDF of a word w is computed as $-log(p(w))$, where p(w) is estimated as the fraction of dictionary 'documents' which contain the word w. Each definition or example in the dictionary is counted as one separate document. At no point are the words stemmed or corrected for case if capitalised.

LESK-DEFINITIONS:
—as LESK, but using only the dictionary definitions, not the dictionary examples. This baseline was included because the HECTOR dictionary has far more examples than most dictionaries, so, where systems assumed more standard dictionaries and did not exploit what was, effectively, a small sense-tagged corpus, LESK-DEFINITIONS would be a more salient baseline.

LESK-CORPUS:
—as LESK, but also considers the tagged training data for words where it is available, so can be compared with supervised systems. For each word w in the sentence containing the test item, this baseline not only tests whether w occurs in the dictionary entry for each candidate sense, but also whether it appears in the same sentence as one of the instances of that sense in the training corpus. That is, (f) above is replaced with:

```
(f') if w occurs in the definition, example sentences or
training-corpus contexts of s,
```

In this case the IDF weights of words are computed for the words' distribution in both the dictionary and the corpus. Each definition or example in the dictionary is counted as one separate document, and also each set of training-corpus contexts for a sense tag is counted as a single additional document. For sense tags which do not appear in the training corpus, the baseline reverts to the strategy of unsupervised LESK, but with the benefit of corpus-derived inverse document frequency weights for words.

Although LESK-CORPUS does not explicitly represent the relative corpus frequencies of sense tags, it implicitly favours common tags because these have larger context sets, and an arbitrary word in a test-corpus sentence is therefore more likely to occur in the context set of a commoner training-corpus sense tag.

. . . +PHRASE-FILTER:
All of the above are also coupled with a phrase filter designed to scan for multi-word expressions in a very shallow way. The phrase filter uses only the dictionary, an inflected-word-forms recogniser, and some rudimentary knowledge about the ordering of the words in each phrase.

The phrase filter is used in conjunction with the baselines as a preprocessor. It runs first, vetoing all senses for multi-word items if there is no evidence for them in the test instance, and vetoing all senses except those for the appropriate multi-word if evidence for one of the dictionary instances is found.

10.1. Prologue to results

Scores were computed on various subsets of the test data, where each subset is intended to highlight a different aspect of the task. There are subtasks for measuring system performance on particular parts of speech, on words for which no training data is available, and on words tagged by the annotators as proper nouns. However, the items on which individual systems significantly outperform or underperform the average did not correlate strongly with any of these broad subsets, so it was not easy to discern which techniques suited which kinds of words or instances.

Individual items in the dataset are not graded in any way for difficulty. This is a limitation of the evaluation since most systems did not tag the entire dataset but carved out more or less idiosyncratic subsets of it, abstaining from guessing about the remainder. Without difficulty ratings for items, we cannot say whether two systems that tag only part of the data have chosen equally hard subsets, and results may not be comparable. In particular, systems which focus on high-frequency phenomena for which reliable cues are available may benefit from saying nothing about more difficult cases.

The highly skewed distribution of language phenomena, with a few very frequent phenomena and a long tail of rarer ones, also means that systems will primarily be evaluated with respect to their ability

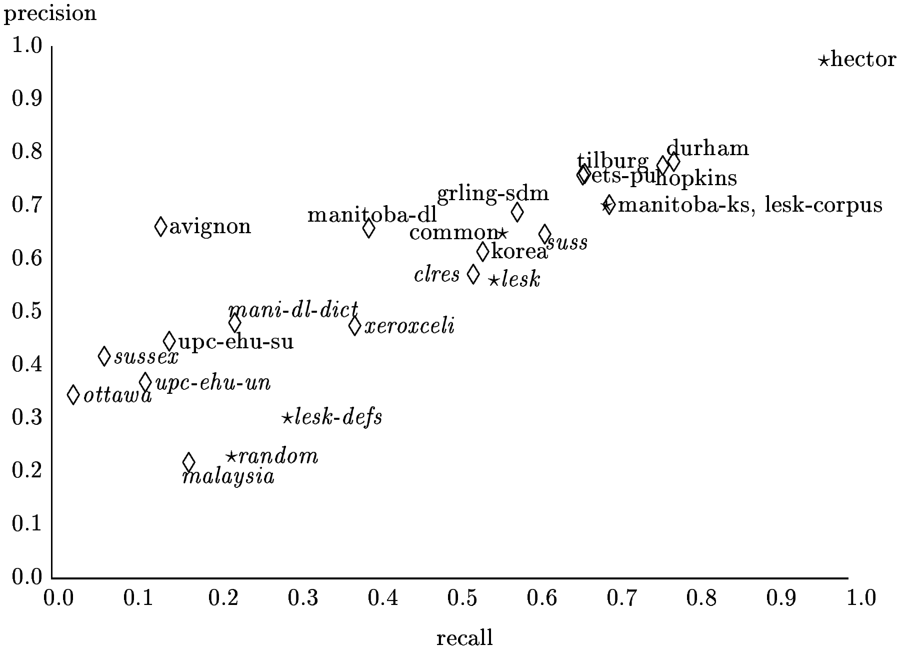

Figure 1. System performance on all test items.

to handle a few common types of problems. Their ability to handle a range of rarer problems will have little impact on their score. Even if a system does not choose to restrict itself to the subset of common cases, there will be little else for it to demonstrate its versatility on.

11. Results for participating systems

The following graphs summarise system performance on several main tasks of the evaluation. Unsupervised systems are in italics, supervised in roman. The human score, HECTOR, corresponds to the annotations made by the lexicographers who initially marked up the test corpus. All the graphs show fine-grained, non-minimal scores.

Five baselines are also provided for comparison: LESK-CORPUS, LESK, LESK-DEFS (all with the phrase filter), COMMONEST and RANDOM. Baselines are bold or italic, according to whether they use the training corpus or not, and have scores marked with stars, where competing systems have diamonds.

Figure 1 demonstrates that the state of the art, for a fine-grained WSD task where there is training data available, is at around 77%: the

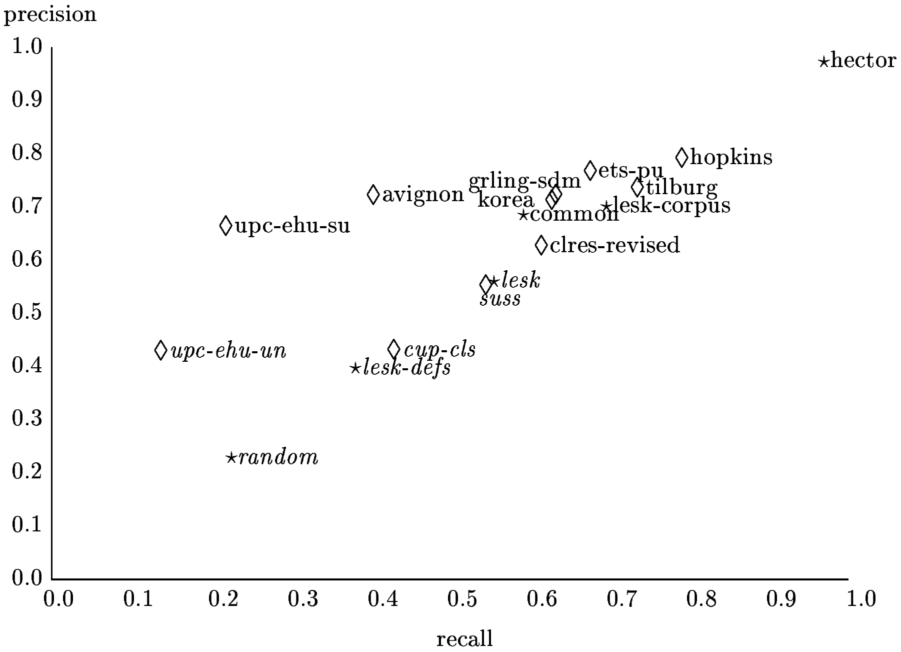

Figure 2. Later-deadline system performance on all test items.

highest scoring system scored 77.1%.[17] Where there is training data available, systems that use it perform substantially better than ones that do not. The Lesk-derived baselines performed well. The majority of systems were outperformed by the best of the baselines for their system-type.

11 systems also returned results by a later deadline. This was mainly to allow further de-bugging, where the rush to meet the pre-workshop deadline had meant the system was still very buggy. Ten of the second-round systems were revised versions of first-round systems and one, `cup-cls`, was a new participant. The highest-scoring of the second-round systems had a marginally higher score (78.1%) than the highest-scoring of the first-round systems. Second-round results are shown in Figure 2.

[17] For the coarse-grained task, the equivalent figures would be 5% higher. The performance of all systems improves under coarse-grained scoring, but in general the relative performance of the systems was not affected (even though some systems had been optimised for the coarse-grain level). The average system precision score on all test items improves from 0.55 to 0.66, or 20%, when scoring is at the coarse-grained instead of the fine-grained level.

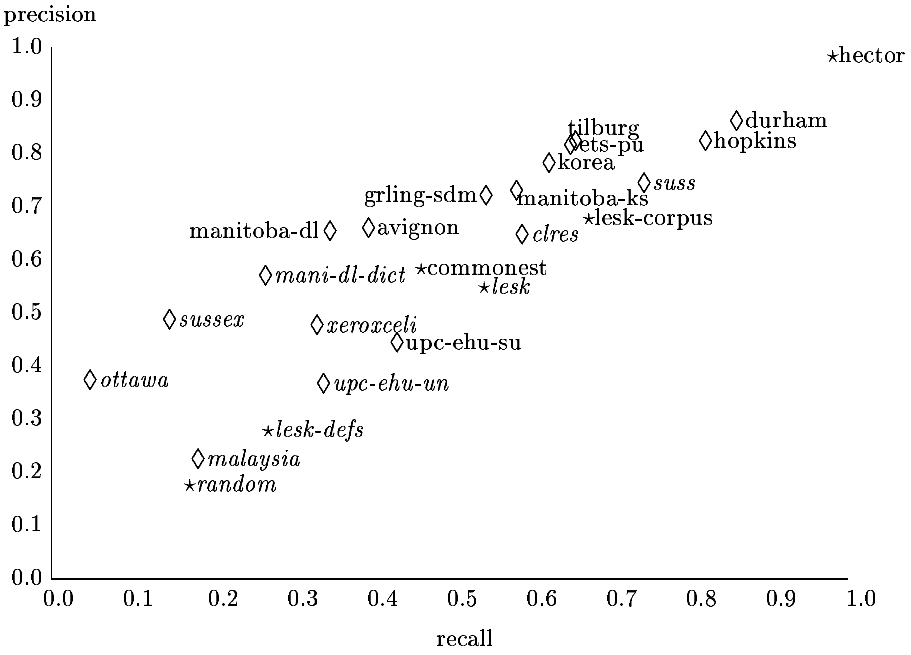

Figure 3. System performance on `nouns` subtask.

Figures 3 and 4 show performance on the nouns and on the verbs. For nouns, the top performance was over 80%; for the verbs, the best systems scored around 70%.[18]

11.1. Tasks with and without training data

Some of the supervised systems (DURHAM, HOPKINS, MANITOBA-DL) were designed to fall back on unsupervised techniques, or to rely on dictionary examples when no corpus training data was available. One might have expected these systems to perform at the same levels as the unsupervised ones for those tasks where there was no training data. But this was not the case. The supervised systems performed better than the unsupervised even for these words.

In general, the systems that attempt both the no-training-data words and the others do better on the no-training-data words. This is a consequence of frequency: corpus data was supplied wherever there was any data left over after the test material was taken out from the

[18] The other two categories, adjectives and **-p** tasks, had top levels between these two.

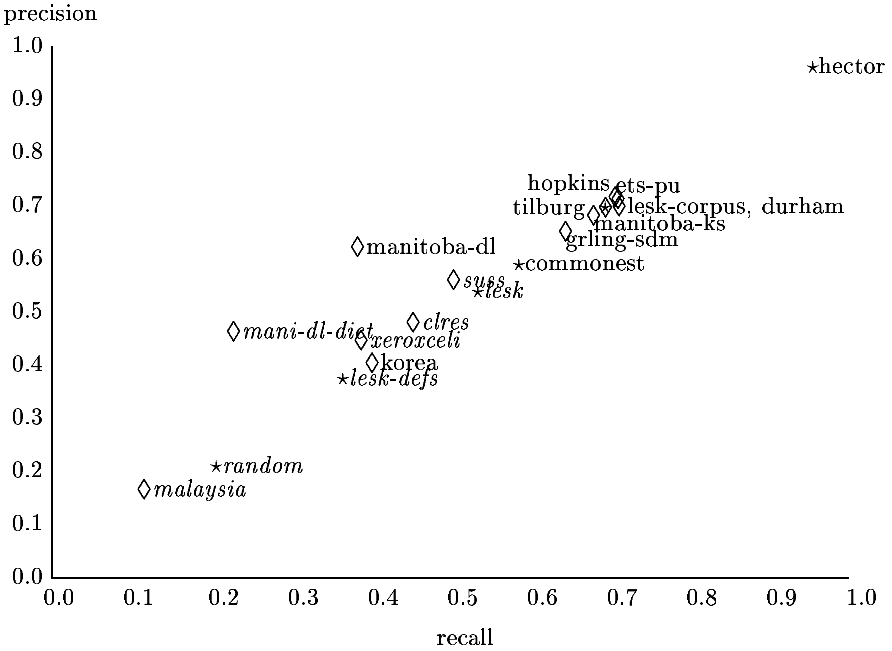

Figure 4. System performance on `verbs` subtask.

HECTOR corpus, so the no-training-data words were the rarer words — and low polysemy is correlated with low frequency: in this case, 7.28 senses per word on average as opposed to 10.79 for words with corpus training data. The entropy is also lower on average: 1.57 versus 1.92 for words with training data.[19] As a result, supervised systems which do not attempt to tag these words are at a disadvantage compared with supervised systems that do somehow manage to tag them.

11.2. Scoring based on reduction of baseline error rate

Participants were free to return guesses for as many or as few of the items as they chose. Hence, participants who, by accident or design, only returned guesses for the easier items may be considered to have inflated scores, and those who have returned guesses for difficult cases, deflated ones. Thus, the SUSSEX system returned guesses for just 879 (10%) of the items in the dataset (just those items where the word to

[19] Entropy is calculated as $-\Sigma(p(x) \cdot log(p(x)))$ where x ranges over all sense tags of a word, and p(x) is the fraction of training occurrences of the word tagged with x.

be tagged was the head of the object-noun-phrase of one of a particular set of high-frequency verbs). The over-all precision of SUSSEX (based on its performance on just these items) is 0.36, as compared to 0.39 for the LESK-DEFINITIONS baseline. However, if we look only at the 879 items for which SUSSEX returned an answer, SUSSEX performed better than the baseline. It so happens that SUSSEX had selected a harder-than-average set of items to return guesses for, and its performance should be seen in that light.

On one large subset of the data, the 2500 items in the verb tasks, none of the systems is capable of achieving more than a 2% improvement over the best baseline's error rate.

11.3. Part-of-speech accuracy

For the **-p** tasks, the input did not provide a part-of-speech specification so the system had, implicitly, to provide one. Most systems guessed part-of-speech correctly over 90% of the time, the two lowest scores being 78% (MANI-DL-DICT) and 87% (MANITOBA-KS). POS-tagging accuracy was not correlated with sense-tagging accuracy.

For most systems, the results relative to baseline are better for **-p** tasks than for **-nva** tasks. For **-nva** tasks, systems and baselines alike can look up the correct part of speech simply by checking the filename suffix. For **-p** tasks, the baselines, unlike the systems, had no POS-tagging module so made many word class errors. For example, TILBURG achieves 13.05% error reduction relative to the LESK-CORPUS baseline. However, much of this is due to the baseline's performance on the indeterminate items, where it makes many more errors simply because it is not equipped with a part-of-speech tagger. If consideration is restricted to the **-nva** task, the error reduction due to TILBURG decreases to 4.52%.

There were a total of 286 items tagged with PROPER in the answer key. These items are always also assigned a dictionary sense tag in addition to PROPER (see section 5.2). Only three systems ever guess PROPER: HOPKINS, TILBURG and ETS-PU.[20] Of these, HOPKINS succeeds in recognising 56.1%, TILBURG 14.3%, and ETS-PU 5.6%. Of the remaining systems, some seem able to distinguish likely proper nouns as they tend to abstain from guessing more often on the PROPER instances.

As discussed in section 5.1, it was possible for a sense tag from the 'wrong' word class to apply, eg., although the SOUND sense for **float** was a sense for verbal **float**, it could be the most salient sense, to be found in the gold standard, for an adjectival instance. Thus the task

[20] PROPER tags do occur in the responses of a couple of other systems, but at most only once or twice per system.

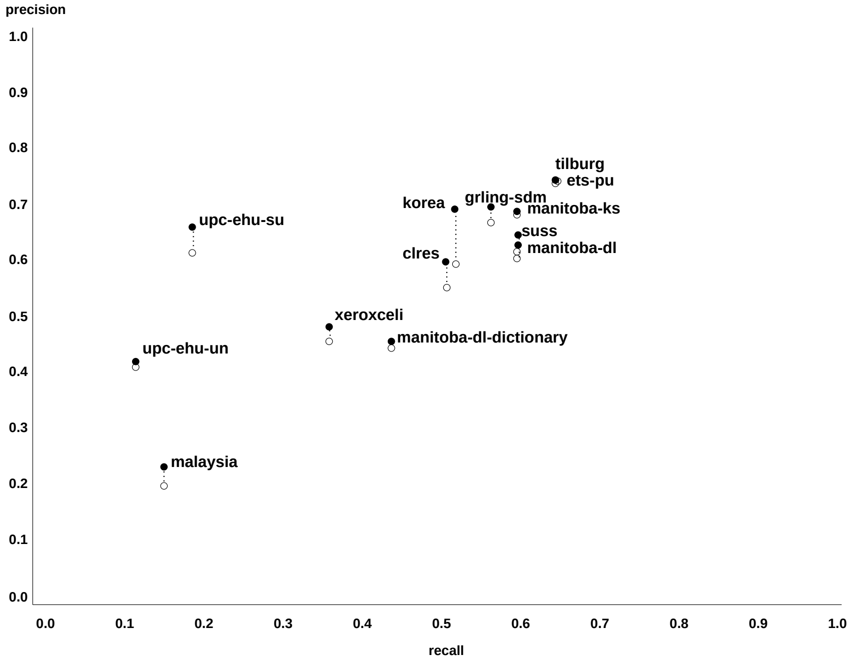

Figure 5. Improvement in system performance when responses are limited to sense tags with a part of speech appropriate to the file type of each test item; unfilled circles show original scores, filled circles, improved ones.

definition permitted any sense tag for any word class for the word (as well as PROPER and UNASS) as possibilities. If that was interpreted as indicating that the **-n**, **-v** or **-a** label on the task imposed no constraint on the sense tags which could apply, then the label provided no, or very little, useful information. In practice, this occurred less than 1% of the time, and systems which only ever guessed ‘right’ word class senses benefited from the simplifying assumption. Systems which did not make this assumption frequently paid heavily, committing 10% of their total errors in this way.[21] Figure 11.3 shows, for those systems, how much their performance improves if we ignore errors which would not have occurred had they heeded the part-of-speech constraint. The shift in precision is accomplished by throwing out any guesses that the system makes in the wrong part of speech. Since all of these were wrong anyway, recall is not affected, but precision increases, sometimes dramatically.

[21] In one case, KOREA, the wrong guesses resulted from a systematic false assumption.

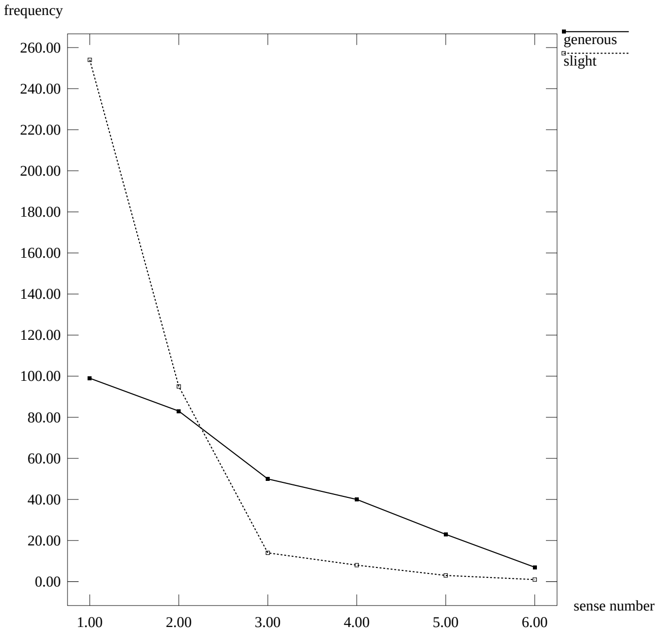

Figure 6. Distribution of sense tags for *generous* versus *slight* in training corpus.

12. Polysemy, entropy, and task difficulty

The distribution of sense tags in the training and evaluation data is highly skewed, with a few very common sense tags and a long tail of rarer ones. This suggests that the distributions of sense tags for individual words in the data will also be quite skewed and that the entropy of these distributions will be fairly low. However, there is substantial variation of entropy across words. For instance, both **generous** and **slight** are adjectives with 6 senses, but the entropy of **slight** is 1.28 while that of **generous** is 2.30. This is because of the unusually even distribution of sense tags for **generous**, as shown in Figure 6 of the training-data distributions for the two adjectives.

Polysemy and entropy often vary together, but not always. As Table V shows, the nouns, on average, had higher polysemy than the verbs but the verbs had higher entropy. For verbs, the corpus instances were spread across the dictionary senses more evenly than for nouns.

Systems tend to do better on the nouns than the verbs, suggesting that entropy is the better measure of the difficulty of the tasks. The correlation between task polysemy and system performance is -0.258. The correlation between entropy and system performance is stronger: -0.510. When considering just the supervised systems, the correlation

Table V. Polysemy and entropy of selected evaluation subtasks

task	average polysemy	average entropy
`eval` (all items)	10.37	1.91
`nouns`	9.16	1.74
`verbs`	7.79	1.86
`adjectives`	6.76	1.66

with entropy is -0.699. Correlation with polysemy for these systems is -0.247.

This might be thought surprising. Where a sense-tag distribution has high entropy, most candidate senses are well-represented in the training corpus, so supervised systems should be able to arrive at good models for all of them and discriminate between them reliably. Against that stand two arguments, one mathematical, one lexicographic. The mathematical one is that low-entropy distributions are often dominated by a single sense, in which case the system can perform well by guessing the dominant sense wherever it does not have good evidence to the contrary. The lexicographic one is this: in deciding what senses to list for a word, lexicographers will only give rarer possibilities the status of a sense where they are quite distinct (Kilgarriff, 1992, chapter 4). Senses which are quite distinct to the lexicographer will tend to be those that are easier for systems to discriminate. At the one end of the spectrum are tasks like **generous-a** where all the meaning distinctions are subtle and overlapping, and the senses tend to be of comparable frequency, giving high entropy for the number of senses. At the other end are tasks like **slight-a** where the sense distinctions are reasonably clear for lexicographers and systems alike, but the rarer senses are far rarer than the dominant one or two, giving low entropy.

The relations between polysemy and precision, and entropy and precision, are depicted in Figures 7 and 8.[22]

There are a few outliers. The two vertical lines on the right of the polysemy graph correspond to the tasks **band-p** (29 senses) and **shake-p** (36 senses). Systems perform quite well on these tasks despite their high polysemy. In the case of **band-p**, this relates to its low entropy (1.75); system performance on *band* is close to system performance

[22] Out-of-candidate-set guesses (for sense tags of the wrong part of speech) have been disregarded in computing the systems' performance on the above graphs, as the inflated polysemy levels, where, eg, all adjectival senses were included as possibilities for a verbal task, would complicate the figure.

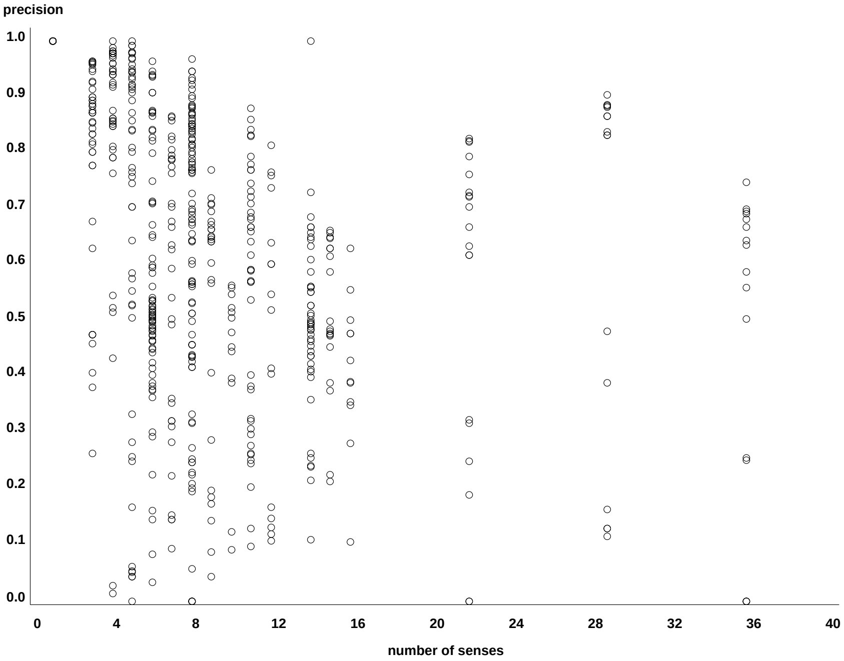

Figure 7. Precision of all systems on words with different numbers of senses.

on other tasks with similar entropy. This in turn relates to the high incidence of compound nominals among the senses of **band**: *big band, band saw, elastic band* etc. These have distinct, unpredictable real-world meanings, so the lexicographer is inclined to treat them as distinct senses even if they are infrequent; for WSD systems, they will be easy to get right.

Shake-p has high entropy (3.69), so the good system performance on this word cannot be explained by the effect of this variable. For **shake**, like **band**, multi-word expressions hold the key. *Shake one's head* is the commonest use of *shake* in the training data, and over 50% of the test items involve some multi-word expression.

13. Pooling the results of multiple systems

Improvements in precision can be achieved by having sets of participating systems vote on which sense tag should be assigned to each test item.[23]

[23] The idea was suggested by Eneko Agirre and David Yarowsky.

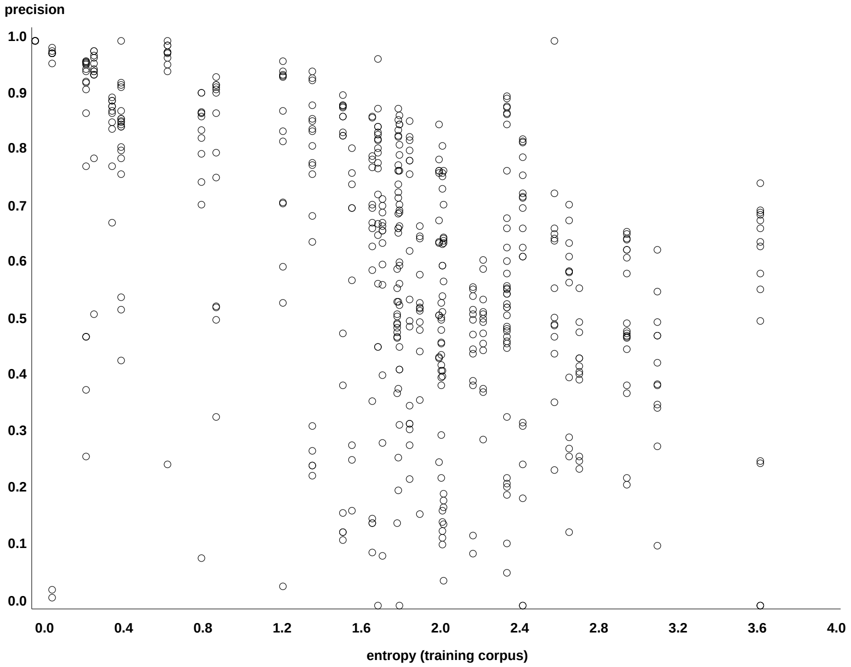

Figure 8. Precision of all systems on words with different entropy measures.

Three voting schemes were explored. UNANIMOUS only assigns a tag if all the systems in the voting pool agree on that tag unequivocally (or abstain from tagging it). ABSOLUTE MAJORITY assigns a tag if one tag gets more of the non-abstaining systems' votes than all the others combined. If no tag gets an absolute majority of votes, no guess is made. WINNER simply guesses the tag or tags that receive more votes than any others. For ABSOLUTE MAJORITY and WINNER, systems which assign weights to multiple sense tags are counted as voting fractionally for each of these sense tags according to the weight they assign them.

The voting schemes were applied to various sets of systems, including: *all* (the complete set of participating systems); *all S* (all the supervised systems); and *best S* (the better half of the supervised systems, as measured by their overall precision).

All the voting schemes gave higher precision than any of their contributing systems. However, all systems agree unanimously on only 3% of items, and even then there are several cases where they do not get the right tag. The agreement between better-performing systems is generally higher than the agreement between systems that do not perform so well.

By combining the best supervised systems in the *best S* voting pool, we achieve 96% precision on a substantial fragment of the dataset (53%). This is comparable to human precision on this task, as measured by the lexicographers' annotations. The recall is of course substantially lower, and the cases that are left out are evidently the more difficult ones. The shallow LESK-PLUS-CORPUS baseline with the PHRASE FILTER attains 86.4% precision on the same subset of the test data, as compared with 49.4% on the remaining test items which the voting pool cannot agree on. The voting pool therefore achieves 66.2% error reduction over the baseline on the fragment of the test data that it tags, as opposed to the 85.1% that one would expect if the items tagged by the voting pool were an arbitrary sample of the test data. But such a high-precision partial annotation, produced automatically, can still be extremely useful. It can serve as a valuable first pass over raw data, and one can anticipate it being used in a variety of ways, including the preparation of gold standard data for future SENSEVALS.

14. Conclusion

English SENSEVAL was an engaging and successful exercise. The strategy developed for the evaluation made evaluation possible and meaningful. Others have worried that WSD cannot be meaningfully evaluated because people so often disagree on what the correct sense is; in the course of the data preparation phase, this ghost was laid to rest, as the human sense-tagging proved to be replicable with a high degree of accuracy.

There was a very high level of interest and engagement with the exercise, with eighteen systems from sixteen research groups participating. Participants were in general grateful that the exercise had been organised, as it enabled them to find out how their system (and its various components) compared with others, in a way that had been near impossible before. It also promoted the coherence of the field through providing a common reference point for evaluation data and methodology.

The exercise identified the state-of-the-art for fine-grained WSD. Where a reasonable quantity of pre-tagged training data was available, the best current systems were accurate 74–78% of the time (where they aimed to tag all instances, ie. maximising recall). It is interesting to note that a number of systems had very similar scores at the top end of the range, and that the LESK-CORPUS baseline, which simply used overlap between words in the training data and test instance, was not far below, at 69%. For systems that did not assume the availability of training data, scores were both lower and more variable. Where training data was available, there has been some convergence on the appropriate methods to use, but where a dictionary is the major source, there has been no such convergence.

System performance correlates more closely with entropy than with polysemy. However there are many outliers and exceptions, and there remains much work to be done in identifying which kinds of words are easy for WSD, and which are difficult.

Limitations of the exercise included the limited amount of context available for each test instance; the small number of words investigated; and, most centrally, uncertainty about the sense inventory that had been selected for the exercise. HECTOR senses may be as valid as those from any other dictionary, but was that good enough? Were they relevant for any NLP task that a WSD module might be useful for? This issue is discussed further in the Introduction and discussion papers in the Special issue.

We believe SENSEVAL has done much to take WSD research forward. We look forward to future SENSEVALs with the continued engagement and co-operation of all researchers in the area.

References

Atkins, S.: 1993, 'Tools for computer-aided corpus lexicography: the Hector project'. *Acta Linguistica Hungarica* **41**, 5–72.

Byrd, R. J., N. Calzolari, M. S. Chodorow, J. L. Klavans, M. S. Neff, and O. A. Rizk: 1987, 'Tools and Methods for Computational Lexicology'. *Computational Linguistics* **13**, 219–240.

CIDE: 1995, 'Cambridge International Dictionary of English'. CUP, Cambridge, England.

Fellbaum, C. (ed.): 1998, *WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press.

Gale, W., K. Church, and D. Yarowsky: 1992, 'Estimating upper and Lower bounds on the performance of word-sense disambiguation programs'. In: *Proceedings, 30th ACL.* pp. 249–156.

Harley, A. and D. Glennon: 1997, 'Combining Different Tests with Additive Weighting and their evaluation'. In: M. Light (ed.): *Tagging Text with Lexical Semantics: Why, What and How?* Washington, pp. 74–78.

Hirschman, L.: 1998, 'The Evolution of Evaluation: Lessons from the Message Understanding Conferences'. *Computer Speech and Language* **12**(4), 281–307.

Jorgensen, J. C.: 1990, 'The Psychological Reality of Word Senses'. *Journal of Psycholinguistic Research* **19**(3), 167–190.

Kilgarriff, A.: 1992, 'Polysemy'. Ph.D. thesis, University of Sussex, CSRP 261, School of Cognitive and Computing Sciences.

Kilgarriff, A.: 1997, 'Evaluating Word Sense Disambiguation Programs: Progress Report'. In: R. Gaizauskas (ed.): *Proc. SALT Workshop on Evaluation in Speech and Language Technology.* Sheffield, pp. 114–120.

Kilgarriff, A.: 1998, 'Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs'. *Computer Speech and Language* **12**(4), 453–472. Special Issue on Evaluation of Speech and Language Technology, edited by R. Gaizauskas.

Lesk, M. E.: 1986, 'Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone'. In: *Proc. 1986 SIGDOC Conference*. Toronto, Canada.

Ng, H. T. and H. B. Lee: 1996, 'Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach'. In: *ACL Proceedings*. Santa Cruz, California, pp. 40–47.

Appendix 1: The CUP-CLS system

The CUP-CLS sense tagger was created at Cambridge University Press with support from the EC funded project ACQUILEX II, and developed further by Cambridge Language Services with support from the DTI/SALT funded project Integrated Language Database, and is fully described in Harley and Glennon (1997). No further modifications have been made to the tagger since that date, and there was no fine-tuning for the HECTOR tags or data. The mapping between CIDE (CIDE, 1995), the dictionary used by the CUP/CLS tagger, and the HECTOR dictionary was done by Guy Jackson to the simple guidelines of noting a map wherever there was an overlap between a CIDE sense and a HECTOR sense. In particular, this meant that many CIDE senses often mapped to one HECTOR sense. This meant that the tagger, which only chooses one CIDE sense for each instance, inevitably tagged many words with multiple HECTOR senses solely because of the mapping. The upper bound for the CIDE mapping (computed as described for WordNet in Section 9) gave figures of 90% attempted and 71% precision. In the evaluation, one of the tags chosen by the CUP/CLS sense tagger after the mapping was correct 64% of the time, i.e. the tagger was definitely wrong 36% of the time. The tagger itself could be improved by a number of measures mentioned in the 1997 paper, in particular by using an external part of speech tagger. (The tagger was not given part of speech information for the evaluation). The mapping could be improved by only mapping the most likely matches not all possible matches, or by mapping to the fine-grained CIDE 'example' level, rather than to the coarser CIDE definition level as now.

References

Harley, A. and D. Glennon: 1997, 'Combining Different Tests with Additive Weighting and their evaluation'. In: M. Light (ed.): *Tagging Text with Lexical Semantics: Why, What and How?* Washington, pp. 74–78.

Appendix 2: The OTTAWA system

The OTTAWA system for word sense disambiguation is part of a larger project that aims to acquire knowledge from technical text semi- automatically. In the absence of hand-coded domain knowledge, the knowledge acquisition tools rely on linguistic knowledge, a cooperating user and general-purpose, publicly available information sources, such as WordNet. For word sense disambiguation, it is possible to use the semantic relationships among nouns in WordNet to compute a measure of semantic similarity of each of the senses of two words. The WSD algorithm attempts to disambiguate nouns by measuring the semantic similarity of senses of words appearing in the same syntactic context: the direct object of a verb. For example, if two nouns appear as direct objects of the same verb, the algorithm measures the similarity of each sense of one noun with each sense of the other noun. The two nouns are disambiguated to the two most similar senses. The algorithm is presented in detail in Li et al. (1995) and Szpakowicz et al. (1996).

References

Li, X., S. Szpakowicz, and S. Matwin: 1995, 'A WordNet-based Algorithm for Word Sense Disambiguation'. In: *Proceedings, IJCAI '95.* Montreal, pp. 1368–1374.

Szpakowicz, S., S. Matwin, and K. Barke: 1996, 'WordNet-based Word Sense Disambiguation that Works for Small Texts'. Technical Report Computer Science TR-96-03, School of Information Technology and Engineering, University of Ottawa.

Appendix 3: The MALAYSIA system

MALAYSIA uses a prescriptive semantic primitive based approach in tagging. Its vocabulary was around 2,000 words for SENSEVAL. The strategy is described in (Wilks et al 1989) and (Guo 1995).

References

Guo, C. -M: 1995, *Constructing a MTD from LDOCE*, Chapt. Part 2, pp. 145–234. Norwood, New Jersey: Ablex.

Wilks, Y., D. Fass, C.-M. Guo, J. McDonald, T. Plate, and B. Slator: 1989, 'A tractable machine dictionary as a resource for computational semantics'. In: B. K. Boguraev and E. J. Briscoe (eds.): *Computational Lexicography for Natural Language Processing*. Harlow: Longman, pp. 193–238.

Appendix 4: HECTOR lexical entries and corpus instances for *generous* and *onion*

GENEROUS: dictionary entry

1 **unstint** (512274)**[adj-qual]** (of a person or an institution) giving willingly more of something, especially money, than is strictly necessary or expected; (of help) given abundantly and willingly

1. *Kodak, one of British athletics' most faithful and generous sponsors, have officially ended their five-year, £5 million backing.* [[= sponsor]]
2. *The British people historically have been extraordinarily generous at disaster giving.* [[subj[person] comp/= at; c/n/giving]]
3. *Grateful thanks to Mr D.S.V. Fosten for his generous help, advice and knowledge freely given.* [[= help]]
4. *It is fashionable to attack doctors for being too liberal in dispensing medication and less than generous with their explanations.* [[= with]]
5. *The US jazz press has been generous in its praise.* [[= in poss nu]]
6. *He was generous with the time he gave to professional organisations.* [[= with time]]

(note=entry is oversplit - WRT)

2 **bigbucks** (512309)**[adj-qual]** (of something monetary) consisting of or representing a large amount of money, sometimes with the implication that the amount is greater than is deserved

1. *The Government is unlikely to be pushed into generous concessions by the rash of public sector disputes.* [[= concession]]
2. *It pays you generous interest on your money.* [[= interest]]
3. *Butler had assembled a complicated financial package which included generous loans to enable the voluntary bodies to build or convert schools for secondary purposes.* [[= [money]]]
4. *Generous offers from News International have helped drive up pay.* [[= offer]]
5. *I can offer you . . . a cheque for the generous sum of £15,000.* [[= [money]]]

3 **kind** (512277)**[adj-qual; often pred]** (of a person or an action) manifesting an inclination to recognize the positive aspects of someone or something, often disinterestedly; (of something that is offered by one person to another) favouring the recipient's interests rather than the giver's

1. *He was always generous to the opposition.* [[= to the opposition]]
2. *His interpretation of my remarks had been generous, often creatively so, making of them something far more brilliant than I had intended.* [[subj/interpretation comp/=]]
3. *This generous desire to show us the best in an author is manifested in his long chapter about Spenser.* [[= desire]]
4. *Some high-minded men believed that the Germans would turn against Hitler if offered generous enough terms.*
5. *The emotions are generous —. altruistic almost —. . . . we feel disturbed personally for other people, for people who have no direct connection with us.* [[subj[emotion] comp/=]]

4 **liberal** (512410)**[adj-qual; often attrib]** leaning toward the positive; liberal

1. *A 25 per cent success rate would be a generous estimate.* [[= estimate]]
2. *Salaries are based on a generous comparison with those paid by the federal civil service of the richest country in the world, the USA.* [[= comparison]]
3. *With the wheels lowered (limiting speed a generous 134 kts) an Apache will settle at 95-100 kts.* [[= [measurement]]]

5 **copious** (512310)**[adj-qual; usu attrib]** (of something that can be quantified) abundant; copious

1. *Serve immediately with generous amounts of fresh Parmesan.* [[= [quantity]]]
2. *In winter protect your cheeks with a generous application of moisturiser.* [[= application]]
3. *Labour spokesmen made generous use of statistics to castigate the government for refusing to spend more money on science.* [[= use]]

6 **spacious** (512275)**[adj-qual; usu attrib]** (of a room or building) large in size; spacious; (of clothing) ample

1. *As if the house were not large enough, there are generous attics stretching right across it, offering another five rooms for expansion.* [[= [room]]]
2. *A generous grill pan large enough to take a family-sized mixed grill* [[= pan]]
3. *A cream crepe dress . . . with generous puffed sleeves and a pleated skirt* [[c/[garment]]]

GENEROUS: corpus instances

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< / > love.
700003
Broderick launches into his reply like a trouper.
‘Oh, it was wonderful, fascinating, a rich experience. He’s a very < *tag* >generous< / > actor and obviously he’s very full.”
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< / > supply of illustrations might have helped the reader place him in the history of 20th-century art.
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< / > and we raised about #200 within the first few hours.”
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 Tues day from folk who may be angling for a favour, a promise or a < *tag* >generous< / > gesture.
700007
Seconds later, airborne missiles whooshed through the air from all directions, apparently aimed at our heads.
It would be < *tag* >generous< / > to call them fireworks, but that implies something decorative, to which one’s response is ‘Aaah”, not ‘Aaagh”.
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< / > 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.
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< / > mood.
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< / > 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.

ONION: dictionary entry

1 **veg** (528347)[**nc, nu**](field=Food) the pungent swollen bulb of a plant, having many concentric skins, and widely used in cooking as a vegetable and flavouring

1. *. . . mutton stew, with potatoes and onions floating in the thickened parsley sauce.*
2. *. . . a finely chopped onion.*
3. *Gently fry the onion and garlic for 5 minutes.*
4. *. . . served with chips, tomatoes, onion rings and side salad.*
5. *. . . french onion soup.*

(kind=cocktail onion, salad onion, Spanish onion, spring onion) (note=cannot separate successfully nu and nc senses)
2 **plant** (528344)[**nc**](field=Botany) the liliaceous plant, Allium cepa, that produces onions, having a short stem and bearing greenish-white flowers; any similar or related plant

1. *When carrots are grown alongside onions, they protect each other from pests.*
2. *Shallots belong to the onion family.*
3. *. . . onion sets*

4. *Allium giganteum is an attractive onion with four feet tall stems topped with dusky purple flowers.*

onion dome

basil (528376)[**nc**](field=Architecture) a bulbous dome on a church, palace, etc

1. *. . . the multicoloured onion domes of St Basil's Cathedral.* [[=]]

(note=typically Russian?)

onion-domed

roofed (528375)[**adj-classif**](field=Architecture) (of a church or other building) having one or more onion domes

1. *Soll is a charming cluster of broad roofed houses and inns sprawling lazily around an onion domed church.* [[=]]

spring onion

spring (528348)[**nc**](field=Botany, Food) a variety of onion that is taken from the ground before the bulb has formed fully, and is typically eaten raw in salads

1. *Garnish with spring onions and radish waterlilies.* [[=]]

ONION: corpus instances

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< / >!" said Maisie.
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* >onions< / > and some Dijon mustard.
700003
With the motor running, slowly add the oil until the mixture is the consistency of a thick mayonnaise.
Stir in the < *tag* >onion< / >, add the salt and pepper or a little more lemon juice if required.
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< / > and parsley and pepper.

700005

Ingredients:

12oz/375g mince 1oz/30ml vegetable or olive oil 2 medium < *tag* >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.

700007

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

Fry, turning frequently to seal the meat.

Add the < *tag* >onion< / >, garlic, carrot, celery and leek and cook for 2 minutes.

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< / > and peppers and cook until soft.

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* >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.

700010

One of the best bulbous plants for drying is Allium albopilosum (christophii). This ornamental < *tag* >onion< / > blooms in June with large globe-shaped flowers up to ten inches in diameter, with small star-shaped silver-lilac flowers.