Bridging the Gap between Dictionary and Thesaurus

Oi Yee Kwong
Computer Laboratory, University of Cambridge
New Museums Site, Cambridge CB2 3QG, U.K.
oyk20@cl.cam.ac.uk

Abstract

This paper presents an algorithm to integrate different lexical resources, through which we hope to overcome the individual inadequacy of the resources, and thus obtain some enriched lexical semantic information for applications such as word sense disambiguation. We used WordNet as a mediator between a conventional dictionary and a thesaurus. Preliminary results support our hypothesised structural relationship, which enables the integration, of the resources. These results also suggest that we can combine the resources to achieve an overall balanced degree of sense discrimination.

1 Introduction

It is generally accepted that applications such as word sense disambiguation (WSD), machine translation (MT) and information retrieval (IR), require a wide range of resources to supply the necessary lexical semantic information. For instance, Calzolari (1988) proposed a lexical database in Italian which has the features of both a dictionary and a thesaurus; and Klavans and Tzoukermann (1995) tried to build a fuller bilingual lexicon by enhancing machine-readable dictionaries with large corpora.

Among the attempts to enrich lexical information, many have been directed to the analysis of dictionary definitions and the transformation of the implicit information to explicit knowledge bases for computational purposes (Amsler, 1981; Calzolari, 1984; Chodorow et al., 1985; Markowitz et al., 1986; Klavans et al., 1990; Vossen and Copestake, 1993). Nonetheless, dictionaries are also infamous of their non-standardised sense granularity, and the taxonomies obtained from definitions are inevitably ad hoc. It would therefore be a good idea if we can unify our lexical semantic knowledge by some existing, and widely exploited, classifications such as the system in Roget's Thesaurus (Roget, 1852), which has remained intact for years and has been used in WSD (Yarowsky, 1992).

While the objective is to integrate different lexical resources, the problem is: how do we reconcile the rich but variable information in dictionary senses with the cruder but more stable taxonomies like those in thesauri?

This work is intended to fill this gap. We use WordNet as a mediator in the process. In the following, we will outline an algorithm to map word senses in a dictionary to semantic classes in some established classification scheme.

2 Inter-relatedness of the Resources

The three lexical resources used in this work are the 1987 revision of Roget's Thesaurus (ROGET) (Kirkpatrick, 1987), the Longman Dictionary of Contemporary English (LDOCE) (Procter, 1978) and WordNet 1.5 (WN) (Miller et al., 1993). Figure 1 shows how word senses are organised in them. As we have mentioned, instead of directly mapping an LDOCE definition to a ROGET class, we bridge the gap with WN, as indicated by the arrows in the figure. Such a route is made feasible by linking the structures in common among the resources.

Words are organised in alphabetical order in LDOCE, as in other conventional dictionaries. The senses are listed after each entry, in the form of text definitions. WN groups words into sets of synonyms ("synsets"), with an optional textual gloss. These synsets form the nodes of a taxonomic hierarchy. In ROGET, each semantic class comes with a number, under which words are first assorted by part of speech and then grouped into paragraphs according to the conveyed idea.

Let us refer to Figure 1 and start from word $x2$ in WN synset X. Since words expressing every aspect of an idea are grouped together in ROGET, we can therefore expect to find not only words in synset X, but also those in the coordinate WN synsets (i.e. M and P, with words m_1, m_2, p_1, p_2, etc.) and the superordinate WN synsets (i.e. C and A, with words c_1, c_2, etc.) in the same ROGET paragraph. In other words, the thesaurus class to which $x2$ belongs should include roughly $X \cup M \cup P \cup C \cup A$. Meanwhile, the LDOCE definition corresponding to the sense of synset X (denoted by D_x) is expected to be similar to the textual gloss of synset X (denoted by $Gl(X)$). In addition, given that it is not unusual for

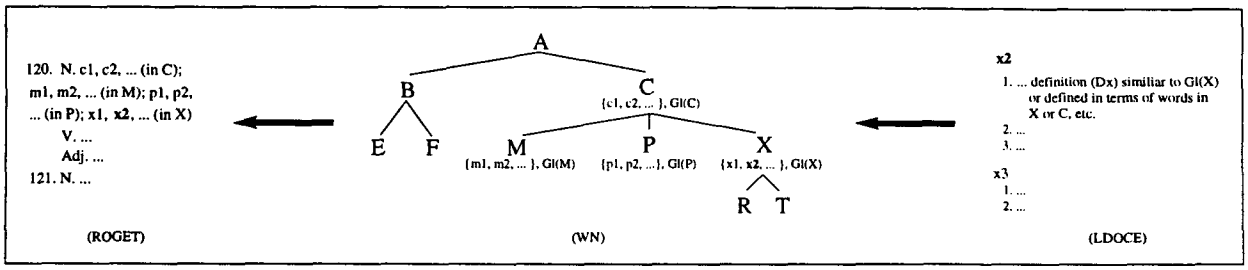

Figure 1: Organisation of word senses in different resources

dictionary definitions to be phrased with synonyms or superordinate terms, we would also expect to find words from X and C, or even A, in the LDOCE definition. That means we believe $D_x \approx Gl(X)$ and $D_x \cap (X \cup C \cup A) \neq \phi$.

3 The Algorithm

The possibility of using statistical methods to assign ROGET category labels to dictionary definitions has been suggested by Yarowsky (1992). Our algorithm offers a systematic way of linking existing resources by defining a mapping chain from LDOCE to ROGET through WN. It is based on shallow processing within the resources themselves, exploiting their inter-relatedness, and does not rely on extensive statistical data. It therefore has an advantage of being immune to any change of sense discrimination with time, since it only depends on the organisation but not the individual entries of the resources. Given a word with part of speech, $W(p)$, the core steps are as follows:

Step 1: From LDOCE, get the sense definitions $D_1, ..., D_t$ under the entry $W(p)$.

Step 2: From WN, find all the synsets $S_n\{w_1, w_2, ...\}$ such that $W(p) \in S_n$. Also collect the corresponding gloss definitions, $Gl(S_n)$, if any, the hypernym synsets $Hyp(S_n)$, and the coordinate synsets $Co(S_n)$.

Step 3: Compute a similarity score matrix $\mathcal{A}$ for the LDOCE senses and the WN synsets. A similarity score $\mathcal{A}(i,j)$ is computed for the i^{th} LDOCE sense and the j^{th} WN synset using a weighted sum of the overlaps between the LDOCE sense and the WN synset, hypernyms, and gloss respectively, that is

$$\mathcal{A}(i,j) = a_1|D_i \cap S_j| + a_2|D_i \cap Hyp(S_j)| + a_3|D_i \cap Gl(S_j)|$$

For our tests, we tried setting $a_1 = 3$, $a_2 = 5$ and $a_3 = 2$ to reveal the relative significance of finding a synonym, a hypernym, and any word in the textual gloss respectively in the dictionary definition.

Step 4: From ROGET, find all paragraphs $P_m\{w_1, w_2, ...\}$ such that $W(p) \in P_m$.

Step 5: Compute a similarity score matrix $\mathcal{B}$ for the WN synsets and the ROGET classes. A similarity score $\mathcal{B}(j,k)$ is computed for the j^{th} WN synset (taking the synset itself, the hypernyms, and the coordinate terms) and the k^{th} ROGET class, according to the following:

$$\mathcal{B}(j,k) = b_1|S_j \cap P_k| + b_2|Hyp(S_j) \cap P_k| + b_3|Co(S_j) \cap P_k|$$

We have set $b_1 = b_2 = b_3 = 1$. Since a ROGET class contains words expressing every aspect of the same idea, it should be equally likely to find synonyms, hypernyms and coordinate terms in common.

Step 6: For $i = 1$ to t (i.e. each LDOCE sense), find $max(\mathcal{A}(i,j))$ from matrix $\mathcal{A}$. Then trace from matrix $\mathcal{B}$ the j^{th} row and find $max(\mathcal{B}(j,k))$. The i^{th} LDOCE sense should finally be mapped to the ROGET class to which P_k belongs.

We have made an operational assumption about the analysis of definitions. We did not attempt to parse definitions to identify genus terms but simply approximated this by using the weights a_1, a_2 and a_3 in Step 3. Considering that words are often defined in terms of superordinates and slightly less often by synonyms, we assign numerical weights in the order $a_2 > a_1 > a_3$. We are also aware that definitions can take other forms which may involve part-of relations, membership, and so on, though we did not deal with them in this study.

4 Testing and Results

The algorithm was tested on 12 nouns, listed in Table 1 with the number of senses in the various lexical resources.

The various types of possible mapping errors are summarised in Table 2. *Incorrectly Mapped* and *Unmapped-a* are both "misses", whereas *Forced Error* and *Unmapped-b* are both "false alarms".

The performance of the three parts of mapping is shown in Table 3. The "carry-over error" is only

Word	R	W	L	Word	R	W	L
Country	3	4	5	Matter	8	5	7
Water	9	8	8	System	6	8	5
School	3	6	7	Interest	14	8	6
Room	3	4	5	Voice	4	8	9
Money	1	3	2	State	7	5	6
Girl	4	5	5	Company	10	8	9

Table 1: The 12 nouns used in testing

Target Exists	**Mapping Outcome**	
	Wrong Match	No Match
Yes	*Incorrectly Mapped*	*Unmapped-a*
No	*Forced Error*	*Unmapped-b*

Table 2: Different types of errors

applicable to the last stage, L →R, and it refers to cases where the final answer is wrong as a result of a faulty outcome from the first stage (L →W).

	L → W	W → R	L → R
Accurately Mapped	68.9%	75.0%	55.4%
Incorrectly Mapped	12.2%	1.4%	4.1%
Unmapped-a	2.7%	6.9%	13.5%
Unmapped-b	13.5%	5.6%	16.2%
Forced Error	2.7%	11.1%	-
Carry-over Error	-	-	10.8%

Table 3: Performance of the algorithm

5 Discussion

Overall, the *Accurately Mapped* figures support our hypothesis that conventional dictionaries and thesauri can be related through WordNet. Looking at the unsuccessful cases, we see that there are relatively more "false alarms" than "misses", showing that errors mostly arise from the inadequacy of individual resources because there are no targets rather than from partial failures of the process. Moreover, the number of "misses" can possibly be reduced if more definition patterns are considered.

Clearly the successful mappings are influenced by the fineness of the sense discrimination in the resources. How finely they are distinguished can be inferred from the similarity score matrices. Reading the matrices row-wise shows how vaguely a certain sense is defined, whereas reading them column-wise reveals how polysemous a word is.

While the links resulting from the algorithm can be right or wrong, there were some senses of the test words which appeared in one resource but had no counterpart in the others, i.e. they were not attached to any links. Thus 18.9% of the LDOCE senses, 11.1% of the WN synsets and 58.1% of the ROGET classes were among these unattached senses. Though this implies the insufficiency of using only one single resource in any application, it also suggests there is additional information we can use to overcome the inadequacy of individual resources. For example, we may take the senses from one resource and complement them with the unattached senses from the other two, thus resulting in a more complete but not redundant sense discrimination.

6 Future Work

This study can be extended in at least two paths. One is to focus on the generality of the algorithm by testing it on a bigger variety of words, and the other on its practical value by applying the resultant lexical information in some real applications and checking the effect of using multiple resources. It is also desirable to explore definition parsing to see if mapping results will be improved.

References

R. Amsler. 1981. A taxonomy for English nouns and verbs. In *Proceedings of ACL '81*, pages 133–138.

N. Calzolari. 1984. Detecting patterns in a lexical data base. In *Proceedings of COLING-84*, pages 170–173.

N. Calzolari. 1988. The dictionary and the thesaurus can be combined. In M.W. Evens, editor, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University Press.

M.S. Chodorow, R.J. Byrd, and G.E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of ACL '85*, pages 299–304.

B. Kirkpatrick. 1987. *Roget's Thesaurus of English Words and Phrases*. Penguin Books.

J. Klavans and E. Tzoukermann. 1995. Combining corpus and machine-readable dictionary data for building bilingual lexicons. *Machine Translation*, 10:185–218.

J. Klavans, M. Chodorow, and N. Wacholder. 1990. From dictionary to knowledge base via taxonomy. In *Proceedings of the Sixth Conference of the University of Waterloo*, Canada. Centre for the New Oxford English dictionary and Text Research: Electronic Text Research.

J. Markowitz, T. Ahlswede, and M. Evens. 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of ACL '86*, pages 112–119.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An online lexical database. *Five Papers on WordNet.*

P. Procter. 1978. *Longman Dictionary of Contemporary English.* Longman Group Ltd.

P.M. Roget. 1852. *Roget's Thesaurus of English Words and Phrases.* Penguin Books.

P. Vossen and A. Copestake. 1993. Untangling definition structure into knowledge representation. In T. Briscoe, A. Copestake, and V. de Paiva, editors, *Inheritance, Defaults and the Lexicon.* Cambridge University Press.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France.