Topical Clustering of MRD Senses Based on Information Retrieval Techniques

Jen Nan Chen*
National Tsing Hua University

Jason S. Chang*
National Tsing Hua University

This paper describes a heuristic approach capable of automatically clustering senses in a machine-readable dictionary (MRD). Including these clusters in the MRD-based lexical database offers several positive benefits for word sense disambiguation (WSD). First, the clusters can be used as a coarser sense division, so unnecessarily fine sense distinction can be avoided. The clustered entries in the MRD can also be used as materials for supervised training to develop a WSD system. Furthermore, if the algorithm is run on several MRDs, the clusters also provide a means of linking different senses across multiple MRDs to create an integrated lexical database. An implementation of the method for clustering definition sentences in the Longman Dictionary of Contemporary English (LDOCE) is described. To this end, the topical word lists and topical cross-references in the Longman Lexicon of Contemporary English (LLOCE) are used. Nearly half of the senses in the LDOCE can be linked precisely to a relevant LLOCE topic using a simple heuristic. With the definitions of senses linked to the same topic viewed as a document, topical clustering of the MRD senses bears a striking resemblance to retrieval of relevant documents for a given query in information retrieval (IR) research. Relatively well-established IR techniques of weighting terms and ranking document relevancy are applied to find the topical clusters that are most relevant to the definition of each MRD sense. Finally, we describe an implemented version of the algorithms for the LDOCE and the LLOCE and assess the performance of the proposed approach in a series of experiments and evaluations.

1. Introduction

Word sense disambiguation (WSD) has been found useful in many natural language processing (NLP) applications, including information retrieval (Krovetz and Croft 1992; McRoy 1992), machine translation (Brown et al. 1991; Dagan, Itai, and Schwall 1991; Dagan and Itai 1994), and speech synthesis (Yarowsky 1992). WSD has received increasing attention in recent literature on computational linguistics (Lesk 1986; Schütze 1992; Gale, Church, and Yarowsky 1992; Yarowsky 1992, 1995; Bruce and Wiebe 1995; Luk 1995; Ng and Lee 1996; Chang et al. 1996). Given a polysemous word in running text, the task of WSD involves examining contextual information to determine the intended sense from a set of predetermined candidates. It is a nontrivial task to divide the senses of a word and determine this set, for word sense is an abstract concept frequently based on subjective and subtle distinctions in topic, register, dialect, collocation, part of speech, and valency (McRoy 1992). Various approaches to word sense division have been proposed in the literature on WSD, including (1) sense numbers in every-day dictionaries (Lesk 1986; Cowie, Guthrie, and Guthrie 1992), (2) automatic or hand-crafted clusters of dictionary senses (Dolan 1994; Bruce and Wiebe 1995; Luk

* Department of Computer Science, National Tsing Hua University, Hsinchu 30043, Taiwan, ROC. E-mail: dr818314@cs.nthu.edu.tw; jschang@cs.nthu.edu.tw.

1995), (3) thesaurus categories (Yarowsky 1992; Chen and Chang 1994), (4) translation in another language (Gale, Church, and Yarowsky 1992; Dagan, Itai, and Schwall 1991; Dagan and Itai 1994), (5) automatically induced clusters with sublexical representation (Schütze 1992), and (6) hand-crafted lexicons (McRoy 1992).

This paper is motivated by the observation that directly using dictionary senses for sense division offers several advantages. Sense distinction according to a dictionary is readily available from machine-readable dictionaries (MRDs) such as the *Longman Dictionary of Contemporary English* (LDOCE) (Proctor 1978). A dictionary such as the LDOCE has broad coverage of word senses, useful for WSD. Furthermore, indicative words and concepts for each sense are directly available in numbered definitions and examples. Lesk (1986) describes the first MRD-based WSD method that relies on the extent of overlap between words in a dictionary definition and words in the local context of the word to be disambiguated. The author reports that WSD performance ranges from 50% to 70% and his method works well for senses strongly associated with specific collocations, such as *ice-cream cone* and *pine cone.*

Unfortunately, using MRDs as the knowledge source for sense division and disambiguation leads to some problems. Zernik (1992) notes that the dictionary dichotomy of senses is inadequate for WSD, because it is defined along grammatical, not semantic, lines. Furthermore, as pointed out in Dolan (1994), the sense division in an MRD is frequently too fine-grained for the purpose of WSD. A WSD system based on dictionary senses often faces unnecessary and difficult "forced-choices." Dolan proposes a heuristic algorithm for forming unlabeled clusters of closely related senses in the LDOCE to eliminate distinctions that are unnecessarily fine for WSD. Regrettably, the proposed algorithm was only described in a few examples and was not developed further. Lacking an automatic method, recent WSD works (Bruce and Wiebe 1995; Luk 1995; Yarowsky 1995) still resort to human intervention to identify and group closely related senses in an MRD.

Using thesaurus categories directly as a coarse sense division may seem to be a viable alternative (Yarowsky 1995). However, typical thesauri, such as *Roget's Thesaurus* (1987), suffer sense gaps and, occasionally, are too fine-grained. Yarowsky (1992) reports that there are uses not listed in *Roget's* for 3 of 12 nouns in his WSD study, while uses which a native speaker might consider as a single sense are often encoded in several *Roget's* categories.

As an alternative approach to word sense division, this paper presents an algorithm capable of automatically clustering senses in an MRD based on topical information in a thesaurus. We refer to the algorithm as *TopSense* (*Top*ical clustering of *Sense*s). The current implementation of *TopSense* uses the topical information in the *Longman Lexicon of Contemporary English* (LLOCE) (McArthur 1992) to cluster LDOCE senses. The method makes use of none of the idiosyncratic information in either the LLOCE or the LDOCE. Therefore, the *TopSense* algorithm is quite general and is expected to produce comparable results for other MRDs and thesauri. *TopSense* is tested on 20 words extensively investigated in recent WSD literature (Schütze 1992; Yarowsky 1992; Luk 1995). According to the experimental results, the automatically derived topical clusters can be used to good effect without any human intervention as a coarse sense division for WSD.

The rest of the paper is organized as follows. Section 2 starts out with a description of the MRDs and thesauri used in the computational lexicography and WSD literature, followed by some observations to justify the topic-based approach to word sense division. Section 3 describes the *LinkSense* algorithm for linking senses between an MRD and a thesaurus. Section 4 shows how the *TopSense* algorithm based on the IR model may be used to cluster the senses in an MRD. Examples are given in both

Sections 3 and 4 to illustrate how the algorithms work. Section 4 also describes an implementation of the algorithms for the LDOCE and the LLOCE and reports the evaluation results for both algorithms based on a 20-word test set. Section 5 analyzes the experimental results to demonstrate the strengths and limitations of the method. The implication of *TopSense* to WSD and other issues related to lexical semantics are also touched upon. Section 6 compares the proposed method with other approaches in the computational linguistics literature. Finally, conclusions are made and directions for further research are pointed out in Section 7.

2. Word Senses in Machine-Readable Dictionaries and Thesauri

In this section, we look at two knowledge sources of word sense division, which are currently widely available, namely, the dictionary and the thesaurus. A good-sized dictionary usually has a large vocabulary and broad coverage of word senses, both of which are useful for WSD. However, a dictionary's division of senses for a given word is often too fine for the task of WSD. On the other hand, a thesaurus organizes word senses into a fixed set of coarse semantic categories, making it more appropriate for WSD. However, thesauri tend to have a smaller vocabulary and a narrower coverage of word senses. To get the best of both worlds of dictionary and thesaurus, we propose to cluster MRD definitions to yield a broad-coverage sense division with the granularity of a thesaurus. Therefore, a short description of MRDs and thesauri is in order.

2.1 Fine-Grained Senses in an MRD

Interest in MRD-based research has increased over the years; in particular, the LDOCE and *Webster's Seventh New Collegiate Dictionary* (W7) (1967) have drawn much attention. Much of the MRD-based research has focused on the analysis and exploitation of the sense definitions in MRDs (Amsler 1984a, 1984b, 1987; Alshawi 1987; Alshawi, Boguraev, and Carter 1989; Vossen, Meijs, and denBroeder 1989). In these works, the definitions are analyzed using either a parser (Montemagni and Vanderwende 1992) or a pattern matcher (Ahlswede and Evens 1988) into semantic relations. These relations are then used for various tasks, ranging from the interpretation of a noun sequence (Vanderwende 1994) or a prepositional phrase (Ravin 1990), to resolving structural ambiguity (Jenson and Binot 1987), to merging dictionary senses for WSD (Dolan 1994). Besides the definition itself, there is an abundance of information listed in a dictionary entry, including part of speech, subcategory, examples, collocations, and typical arguments, which is potentially useful for WSD. In this regard, the LDOCE is particularly appropriate since it uses a reduced, controlled vocabulary of some 2,000 words to define over 60,000 word senses representing a comprehensive vocabulary and broad coverage of word senses.

It is arguable that the dictionary division of senses for a given word is too fine-grained, thus inadequate for WSD. For instance, it might not always be necessary or easy to distinguish between two LDOCE senses **bank.1.n.1** (*river bank*) and **bank.1.n.5** (*sandbank*) shown in Table 1. Hence, dictionary senses can be used to good effect for WSD only if such closely related senses are merged and treated as one. There is more than one way to merge dictionary senses. In the following sections, we describe one such approach, under which MRD senses are merged according to the sense granularity of a typical thesaurus.

2.2 Coarse Senses in Thesauri: WordNet, *Roget's*, and LLOCE

One of the most potentially valuable aspects of the thesaurus, as a knowledge source for word sense division, is the organization of word senses into a limited number of

Table 1
The sense entries for *bank* in the LDOCE.

Sense ID	Sense Entries
bank.1.n.1	land along the side of a river, lake, etc.
bank.1.n.2	earth which is heaped up in a field or garden, often making a border or division.
bank.1.n.3	a mass of snow, clouds, mud, etc.
bank.1.n.4	a slope made at bends in a road or race-track, so that they are safer for cars to go round.
bank.1.n.5	= SANDBANK. (a high underwater bank of sand in a river, harbour, etc.)
bank.2.v.1	(of a car or aircraft) to move with one side higher than the other, esp. when making a turn
bank.3.n.1	a row, esp. of OARs in an ancient boat or KEYs on a TYPEWRITER.
bank.4.n.1	a place in which money is kept and paid out on demand, and where related activities go on.
bank.4.n.2	(usu. in comb.) a place where something is held ready for use, esp. ORGANIC products of human origin for medical use.
bank.4.n.3	(a person who keeps) a supply of money or pieces for payment or use in a game of chance.
bank.5.v.1	to put or keep (money) in a bank.
bank.5.v.2	[esp. with] to keep one's money (esp. in the stated bank)

Note: Sense ID = Root + Homonym No. + Part-of-speech + Sense No.

Table 2
Roget's semantic classes and categories.

Class	Categories	Gloss for Classes
A	1–182	Abstract relations
B	183–318	Space
C	319–446	Matter
D	447–594	Intellect: the exercise of the mind
E	595–816	Volition: the exercise of the will
F	817–990	Emotion, religion and morality

coarse semantic categories. We briefly describe the on-line thesauri, WordNet (Miller et al. 1993), *Roget's Thesaurus*, and LLOCE, which have been used as word sense divisions in the computational linguistics literature. WordNet is organized as a set of hierarchical, conceptual taxonomies of nouns, verbs, adjectives, and adverbs called **synsets**. The synsets are too fine-grained from the WSD perspective; WordNet contains 24,825 noun synsets for 32,264 distinct nouns with a total of 43,136 senses in its noun taxonomy alone. It would be difficult to acquire WSD knowledge for making such fine distinctions even from a substantial body of training materials.

Roget's Thesaurus arranges words in a three-layer hierarchy and organizes over 30,000 distinct words into some 1,000 categories on the bottom layer. These categories are divided into 39 middle-layer sections that are further organized as 6 top-layer classes. Each category is given a three-digit reference code. To make the hierarchical structure explicit, an uppercase letter from *A* to *F* is added to the reference code to denote the top-layer class for each category, as indicated in Table 2. Similarly, the middle layer is denoted with a lowercase reference letter. The sections related to class **B** (*Space*) are shown in Table 3. Therefore, the reference code for each category is denoted by an uppercase class letter, a lowercase section letter, and a three-digit category number.

Table 3
Sections related to the *Space* class in *Roget's*.

Class	Sections		Gloss of Section	Examples
B	183–194	a	Space in general	surface, heavens, room, kitchen, abode
B	195–242	b	Dimensions	weight, proximity, clothes, wear, hall
B	243–264	c	Form	idea, distortion, flat, plug, yawn, subway
B	265–318	d	Motion	rocket, transposition, carrier, entrance

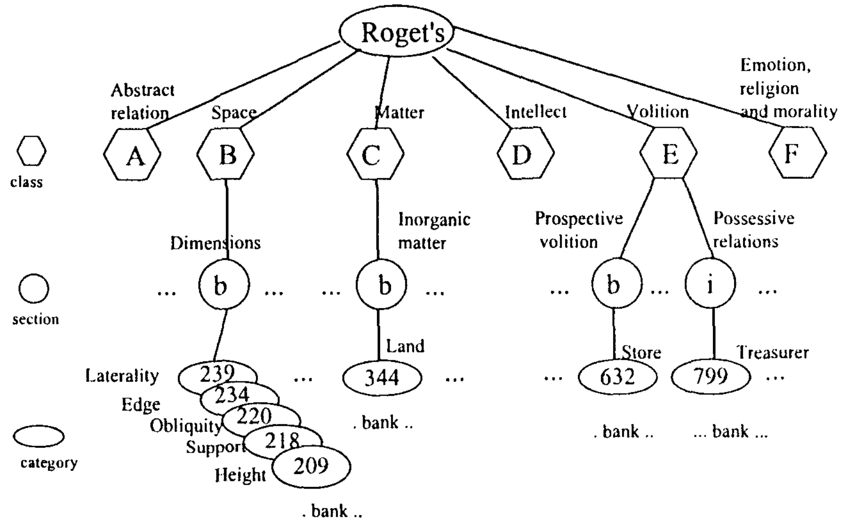

Figure 1
Roget's categorization scheme.

For instance, the word *bank* listed under Category **209** in *Roget's* will be prefixed an additional letter **B** to denote the class *Space*, plus a lowercase letter **b** to denote the section *Dimensions*; the reference code **209** is replaced with **Bb209**. Figure 1 shows the information for the word *bank* in *Roget's*.

WordNet and *Roget's* to a lesser degree present word senses that are too fine-grained for WSD. Often, uses that a native speaker might consider as a single sense are encoded in several *Roget's* categories or WordNet synsets. For instance, a single LDOCE sense **bank.4.n.1** shown in Table 1 corresponds to two WordNet synsets *Depository financial institution* and *Bank building* and two *Roget's* categories, **799** (*Treasurer*) and **784** (*Lending*). Similarly, the *Roget's* lists two categories **234** (*Edge*) and **344** (*Land*) for a concept treated as one word sense, **bank.1.n.1** in the LDOCE. Table 4 provides further details.

The LLOCE is a hierarchical thesaurus that organizes word senses primarily according to subject matter. The LLOCE contains over 23,000 different senses for some 15,000 distinct words. The coarser senses in LLOCE are organized into approximately 2,500 topical word sets. These sets are divided into 129 topics and these topics are further organized as fourteen subjects. The subjects are denoted with alphabetical reference letters from **A** to **N** (see Table 5). Thus the LLOCE subject, topic, and topical set constitutes a three-level hierarchy, in which each subject contains 7 to 12 topics and each topic contains 10 to 50 sets of related words. Table 6 displays the topics related

Table 4
Comparison of MRD and thesaurus treatment of *bank* senses.

LDOCE	WordNet Sense	*Roget's* Sense	LLOCE Sense
bank.1.n.1	Ridge	234 (Edge)/344 (Land)	Ld099 (River bank)
bank.1.n.2	Ridge	234 (Edge)	—
bank.1.n.3	Array	—	—
bank.1.n.4	Slope	239 (Laterality)	—
bank.1.n.5	Ridge	344 (Land)	Ld099 (River bank)
bank.2.v.1	Tip laterally	239 (Laterality)	Nj295 (To bend)
bank.3.n.1	Array	—	—
bank.4.n.1	Depository/Bank building	799 (Treasurer)/784 (Lending)	Je104 (Finance)
bank.4.n.2	Supply	632 (Storage)	—
bank.4.n.3	Supply	—	—
bank.5.v.1	Deposit	799 (Treasurer)	Je106 (Deposit)
bank.5.v.2	Keep money/Deposit	799 (Treasurer)	Je106 (Deposit)

Table 5
LLOCE subjects and their reference letters.

Subject	Set	Gloss for Subjects
A	**1–158**	Life and living things
B	**1–181**	Body; its function and welfare
C	**1–357**	People and family
D	**1–186**	Buildings, houses, home, clothes, belongings, personal care
E	**1–143**	Food, drink, and farming
F	**1–283**	Feeling, emotions, attitudes, and sensations
G	**1–293**	Thought, communication, language, and grammar
H	**1–252**	Substance, materials, objects, and equipment
I	**1–148**	Arts/Crafts, science/technology, industry/education
J	**1–240**	Numbers, measurement, money, and commerce
K	**1–207**	Entailment, sports and games
L	**1–273**	Space and time
M	**1–225**	Movement, location, travel, and transportation
N	**1–367**	General and abstract terms

to subject L (*Space* and *time*). Each topical set is given a three-digit reference code; however, this code does not explicitly reflect the topic. To make use of the information related to a topic, we have designated a lowercase letter to each topic. Therefore, each set is denoted by an uppercase "subject" letter, a lowercase "topic" letter, and a three-digit "topical set" number. For instance, the word *bank* listed under **L99** in the LLOCE will be given an additional reference letter **d** to denote the topic *Geography*; the reference code **L99** is replaced with **Ld099**. The LLOCE also provides cross-references between sets and topics to indicate various intersense relations not captured within the same topic. For instance, topic **Ld** (*Geography*) has a cross-reference to topic **Me** (*Place*). Figure 2 shows LLOCE's topical classification and cross-references related to the word *bank*.

The LLOCE, and, to a lesser degree, *Roget's*, are based on coarse, topical semantic classes, making them more appropriate for WSD than the finer-grained WordNet synsets. The 129 topics in the LLOCE or 990 categories in *Roget's* appear to be suffi-

Table 6
Topics related to subject *L* in LLOCE.

Subject	Range		Gloss	Examples
L	001–019	a	The universe	sun, moon, star, left, right, etc.
L	020–039	b	Light and color	light, dark, ray, color, white, black, etc.
L	040–079	c	Weather and temperature	weather, sky, rain, snow, rain, ice, etc.
L	080–129	d	Geography	stream, sea, lake, flood, to flow, etc.
L	130–169	e	Time generally	time, history, frequent, permanent, etc.
L	170–199	f	Beginning and ending	start, stop, late, last, etc.
L	200–219	g	Old, new, and young	ancient, modern, future, age, etc.
L	220–249	h	Period/Measure of time	day, night, second, minute, etc.
L	250–273	i	Function words (time)	now, soon, always, ever, after, etc.

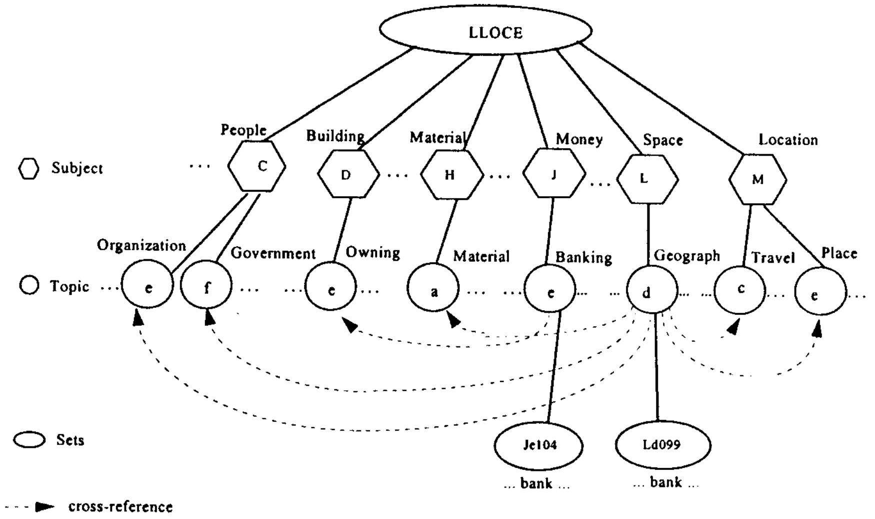

Figure 2
LLOCE's topical organization of word sense.

cient for representing the distinction we would want to make for the task of WSD. *Roget's* has been used as the sense division in two recent WSD works (Yarowsky 1992; Luk 1995) more or less as is, except for a small number of senses added to fill gaps. We contend that a sense division based on the LLOCE topics will offer more or less the same kind of granularity, suitable for WSD. For instance, in Yarowsky (1992), the senses of *star* are divided into three *Roget's* categories, which roughly correspond to five LDOCE *star* senses labeled with LLOCE topics. In the same study, six *Roget's* categories are sufficient to distinguish the senses of *slug*. These six categories correspond to five relevant LLOCE topics. Table 7 provides further details.

2.3 Combining Word Sense Information from an MRD and a Thesaurus

It should be clear by now that combining a dictionary and a thesaurus leads to a broad-coverage sense division with a suitable granularity for WSD. The obvious way to combine the two would be to disambiguate and link a sense definition *D* of a headword *h* in the dictionary to an entry relevant to *D* in the thesaurus. This amounts to a special case of WSD with respect to thesaurus senses. There is no simple solution

Table 7
Roget's and LLOCE classifiers for two sample words.

Word	*Roget's* (three-layer representation)	LLOCE (two-layer representation)
star	**321** (*Universe*)	**La** (*Universe*)
	594 (*Entertainer*)	**Kd** (Drama)
	729 (*Insignia*)	**Jb** (*Mathematics*)
	—	**Dg** (*Personal belongings*)
	—	**Nb** (*Chance*)
slug	**365** (*Animal/Insect*)	**Ag** (*Insect*)
	587 (*Printing*)	**Gd** (*Communication*)
	359 (*Impulse/Impact*)	—
	797 (*Money*)	**Jd** (*Money*)
	723 (*Arms*)	**Hh** (*Weapon*)
	322 (*Weight*)	**Hc** (*Specific substances*)

to the general WSD problem for unrestricted text, but we will show that this special case of disambiguating MRD definitions is significantly easier, for several reasons.

First, the words used in a definition sentence are limited primarily to a small set; in the case of the LDOCE, the controlled vocabulary consists of some 2,000 words. For instance, in the first five LDOCE senses of *bank* shown in Table 1, all defining words are in the controlled vocabulary, except for the word *SANDBANK*, shown in capital letters. Obtaining WSD information for this small set of words obviously is much easier than it would be for a large, open set.

Second, dictionary definitions adhere to rather rigid patterns under which only words with predictable semantic relations show up. A dictionary definition, in general, begins with a **genus term** (that is, conceptual ancestor of the sense), followed by a set of **differentiae** that are words semantically related to the sense to provide the specifics. The semantic relations between the sense, the genus, and differentiae are reflected in what are termed **categorical**, **functional**, and **situational** clusters in McRoy (1992). The semantic relations and clusters have been shown to be very effective knowledge sources for such NLP tasks as WSD (McRoy 1992) and interpretation of noun sequences (Vanderwende 1994). For instance, in the first four definitions of *bank* in Table 1, the genus terms *land, earth, mass,* and *slope* are categorically related to the respective *bank* senses. On the other hand, the differentiae *river, lake, field, garden, bend, road,* and *racetrack* have a *LocationOf* situational relation with *bank*. Other differentiae, *snow, cloud,* and *mud,* are related functionally to **bank.1.n.3** through the *MakeOf* relation.

Third, for the most part these relations are captured implicitly in a typical thesaurus. The LLOCE and *Roget's* conveniently contain information on the relations in the form of word lists under a topic (category) or cross-referencing to other topics. Therefore, an MRD sense definition can be effectively disambiguated based on the word lists and cross-references in a thesaurus. A simple heuristic relying on the similarity between a sense's defining keywords and thesaurus word lists suffices to link an MRD sense to its relevant sense in the thesaurus. For instance, the differentiae (*land, side, river, lake*) of **bank.1.n.1** is sufficiently similar to the word list of **Ld**-topic (*Geography*) to warrant the link between LDOCE sense **bank.1.n.1** and LLOCE sense **bank-Ld099**.

The topics and cross-references of LLOCE in general capture the *Generic/Specific* relation; therefore, a sense definition is often disambiguated through the genus. Thus, the task of linking MRD and thesaurus senses is closely related to the extraction and disambiguation of the genus. For instance, in the above example, linking **bank.1.n.1** to **bank-Ld099** has, as a by-product, the disambiguation of the genus *land* to **land-Ld084** (*Geography*) rather than **land-Ce078** (*Social organization in groups and place*). Details of extraction and disambiguation of the genus can be found in previous works (Guthrie et al. 1990; Klavans, Chodorow, and Wacholder 1990; Copestake 1990; Ageno et al. 1992). Disambiguated genus and differentiae terms can be used to construct a better taxonomy of word senses.

Since the dictionary usually has broader coverage of word senses than the thesaurus, not all MRD senses of a headword *h* correspond to one of *h*'s predefined senses in the thesaurus. For instance, LDOCE sense **bank.1.n.3** (*a mass of cloud, snow, or mud, etc.*) corresponds to LLOCE topic **Hb** (*Object generally*) rather than any of the predefined LLOCE senses for *bank*. Therefore, such entries represent sense gaps in the thesaurus and should be left unlinked. Nevertheless, the linked entries are enough training material for topical clustering of MRD senses, as described in Section 4.

3. Linking an MRD to a Thesaurus

This section describes how to establish a link between an MRD sense and its relevant word sense in a thesaurus, if such a link exists. We start with the preprocessing steps for the sense definition, which are necessary for the algorithm to obtain good results. Then we describe the linking algorithm step by step. Finally, we show illustrative examples to give some idea how the proposed algorithm works for the LLOCE and *Roget's*.

3.1 Preprocessing Steps

Although only simple words are usually used in sense definitions, most of these words are also highly ambiguous. For instance, the two instances of *lies* listed in the two following LDOCE sense definitions differ in meaning:

couch.2.n.2 a bed-like piece of furniture on which a person *lies* when being examined by a doctor.

lie detector an instrument that is supposed to show when a person is telling *lies*.

Notably, their parts-of-speech are also different. Determining the part of speech of each instance allows us to limit the range of possible meanings. The first instance of *lies* is a verb that means "to be in a flat resting position" or "to tell a lie." On the other hand, the second instance is a nominal with a unique meaning "a false statement purposely made to deceive." By tagging the definition with part-of-speech information, the degree of sense ambiguity in the definition can be reduced, thereby increasing the chance of successful linking.

Part-of-Speech Tagging. Various methods for POS tagging have been proposed in recent years. For simplicity, we adopted the method proposed by Church (1988) to tag definition sentences. Experiments indicated an average error rate for tagging of less than 10%. Tagging errors have limited negative impact, because words in the LLOCE are organized primarily according to topic, not part of speech. The POS information is used to remove function words, as well as to look up words in the LLOCE with matching POS. The part-of-speech preprocessing phase is mandatory for the algorithm to exclude some inappropriate candidates for topics. See Table 8 for some examples of tagged LDOCE definition sentences.

Table 8
Some tagged LDOCE definition sentences for the headword *bank*.

bank.1.n.1	land/n along/prep the/det side/n of/prep a/det river/n ,/, lake/n ,/, etc/adv
bank.1.n.2	earth/n which/det is/v heaped/v up/adv in/prep a/det field/n or/conj garden/n ,/, often/adv making/v a/det border/n or/conj division/n
bank.1.n.3	a/det mass/n of/prep snow/n ,/, clouds/n ,/, mud/n ,/, etc/adv
bank.1.n.4	a/det slope/n made/v at/prep bends/n in/prep a/det road/n or/conj race-track/n ,/, so/conj that/conj they/pron are/v safer/adj for/conj cars/n to/* go/v round/adv

Table 9
Examples of keywords extracted from tagged definition sentences.

bank.1.n.1	land/n side/n river/n lake/n
bank.1.n.2	earth/n heap/v field/n garden/n border/n division/n
bank.1.n.3	mass/n snow/n clouds/n mud/n
bank.1.n.4	slope/n bend/n road/n race-track/n carš/n

Removal of Stopwords. In general, function words in the definition are only marginally relevant to the meaning being defined. This is also true of words used in many definitions. For this reason, IR systems commonly exclude stopwords from the process of indexing and query. This also applies to our situation of retrieving topics relevant to the meaning of a sense based on the words in its definition. The list of all the stopwords is specifically designed to remove pronouns, determiners, prepositions, and conjunctions. Table 9 shows that the meaning of some definitions of *bank* is found to be quite intact, even after stopwords are removed.

Calculating Similarity between Definition and Thesaurus Class. When viewing the definition of a headword h as a set of words, it becomes easy to compare and measure their similarity to thesaurus word classes containing h. By word classes, we mean any supersets of synonym sets in a thesaurus that capture the semantic relations and semantic clusters that are effective for disambiguation as described in Section 2.3. The word classes are so chosen that they contain enough words to overlap with the sense definition in question. But each class should not be so big as to cover more than one thesaurus sense for h, blurring the distinction we want to make in the first place. Topics in the LLOCE and categories or sections in *Roget's* are good choices for such classes. Similarity between the defining keywords and a class of words reflects how closely the definition is related to the class. As a simple heuristic, the intended meaning of a dictionary definition D for h is disambiguated in favor of a relevant sense T for h in a thesaurus class C with the highest similarity to D. When such a sense T is found, we say that the dictionary sense D is linked to the thesaurus sense T or that D is linked to the thesaurus class C (containing T.)

For a headword h, let DEF_h denote the definitions of h and let $CLASS_h$ be the word classes in a thesaurus that contain h. For a definition $D \in DEF_h$, our problem amounts to finding $C \in CLASS_h$ that is relevant to D. With these terms, the unweighted Dice coefficient can be adopted to measure similarity between a definition D and a class C as follows:

$$\mathrm{Sim}(D,C) = \frac{\sum_{d \in \mathrm{KEY_D}} 2 \times w_d \times \mathrm{In}(d,C)}{|\mathrm{KEY_D}| + |\mathrm{C}|},$$

where KEY_D = the set of words in definition $D \in DEF_h$, $|KEY_D|$ = number of words in

KEY_D, $C \in CLASS_h$ = a relevant class to h in the thesaurus, $w_k = \frac{1}{\text{degree of ambiguity of } k}$, $In(a, B) = 1$, when $a \in B$, and $In(a, B) = 0$, when $a \notin B$.

The above similarity measure may be improved by taking into consideration specific features of a particular thesaurus. For instance, the cross-reference features in the LLOCE or the intersense relations in *Roget's* are very effective in reflecting semantic relatedness; thus, they should be included in this similarity measure. Let REF_C represent the cross-referenced classes for the word class C in the thesaurus. Thus, we have

$$\mathrm{Sim}'(D, C) = \frac{\sum_{\mathrm{d} \in \mathrm{KEY_D}} 2 \times \mathrm{w_d} \times (\mathrm{In}(\mathrm{d}, \mathrm{C}) + \gamma\ \mathrm{In}(\mathrm{d}, \mathrm{REF_C}))}{|\mathrm{KEY_D}| + |\mathrm{C}| + \gamma|\mathrm{REF_C}|},$$

where γ = relevancy of cross-references to a class[1], and $|REF_C|$ = the number of classes in REF_C.

3.2 The *LinkSense* Algorithm

We sum up the above description and outline the procedure for labeling senses on a dictionary entry as follows:

Algorithm *LinkSense*
Linking fine-grained MRD senses to their relevant thesaurus classes.

Step 1: Given a head word h, read its definition, DEF_h, from the MRD.

Step 2: For each definition D in DEF_h, tag each word in D with POS information.

Step 3: Remove all stop words in D to obtain a list of keyword-POS pairs, KEY_D.

Step 4: Look up the headword h in the thesaurus to obtain $CLASS_h$.

Step 5: Compute $\mathrm{Sim}(D, C)$ for all $C \in CLASS_h$.

Step 6: Link D to C such that $\mathrm{Sim}(D, C)$ is the largest and $\mathrm{Sim}(D, C)$ is greater than a preset threshold, θ.

3.3 Illustrative Examples: Linking LDOCE to LLOCE and *Roget's*

Two examples are given in this subsection to illustrate how *LinkSense* works to establish linkage between a typical dictionary and thesaurus. Example 1 shows, step by step, how *LinkSense* links up an LDOCE sense, **interest.1.n.2** (*a share in a company business etc.*) with the relevant LLOCE sense **interest-Je** (*Banking*).[2] Example 2 is intended to show that *LinkSense* is quite general and applies to thesauri other than the LLOCE. The same LDOCE senses will be shown to links to a relevant *Roget's* sense **interest-Ei** (*Possessive relation*).

Example 1
Linking an LDOCE sense **interest.1.n.3** to its relevant LLOCE sense.

Step 1: D = *"a share in a company, business, etc."*

Step 2: POS_D = {a/det, share/n, in/prep, a/det, company/n, business/n, etc./adv}

[1] For simplicity, the parameter γ is set to 1 in our experiment.

[2] **Je** represents the class of words related to the topic of *Banking, Wealth, and Investment* listed under LLOCE topical sets **Je100** through **127**. The reference code **e** is added in accordance with the coding scheme described in Section 2.2.

Step 3: $KEY_{\mathrm{D}} = \{\text{share/n, company/n, business/n}\}$, $|KEY_{\mathrm{D}}| = 3$.

Step 4: Using LLOCE topics as word classes in *LinkSense*, we have

$$CLASS_{\text{interest}} = \{\mathbf{Fj}\ (\textit{Excitement}),\ \mathbf{Fb}\ (\textit{Liking}),\ \mathbf{Je}\ (\textit{Banking}),\ \mathbf{Ka}\ (\textit{Entertainment})\},$$

The LLOCE lists the following cross references relevant to $CLASS_{\text{interest}}$:

$$\begin{aligned}
\mathrm{REF}_{\mathrm{Fj}} &= \{\mathbf{Ka}\ (\textit{Entertainment}),\ \mathbf{Kb}\ (\textit{Music and related activity}), \ldots,\\
&\quad\ \mathbf{Kh}\ (\textit{Outdoor games})\},\\
\mathrm{REF}_{\mathrm{Fb}} &= \{\mathbf{Cc}\ (\textit{Friendship})\}, REF_{\mathrm{Je}} = \{\mathbf{De}\ (\textit{Getting and giving})\},\\
\mathrm{REF}_{\mathrm{Ka}} &= \{\mathbf{Fj}\ (\textit{Excitement})\},\ |\mathbf{Fj}| = 1,\ |\mathbf{Fb}| = 1,\ |\mathbf{Je}| = 1,\ |\mathbf{Ka}| = 1,\\
&\quad |\mathrm{REF}_{\mathbf{Fj}}| = 8,\ |\mathrm{REF}_{\mathbf{Fb}}| = 1,\ |\mathrm{REF}_{\mathbf{Je}}| = 1,\ |\mathrm{REF}_{\mathbf{Ka}}| = 1.
\end{aligned}$$

All three keywords appear in three different topics but only the following classes are relevant to $CLASS_{\text{interest}}$: **De** (*share*), **Je** (*share*), **Cc** (*company*) Thus, we have

$$\begin{aligned}
&\mathrm{In}(\text{share},\ \mathbf{De}) = 1,\ \mathrm{In}(\text{share}, \mathbf{Je}) = 1,\ \mathrm{In}(\text{company}, \mathbf{Cc}) = 1.\\
&w_{\text{share/n}} = w_{\text{company/n}} = w_{\text{business/n}} = 1/3.
\end{aligned}$$

Step 5: Similarity values are calculated as follows:

$$\begin{aligned}
\mathrm{Sim}'(D, \mathbf{Je}) &= \frac{2 \times w_{\text{share}} \times (In(\text{share}, \mathrm{Je}) + In(\text{share}, REF_{\mathrm{Je}}))}{|\{\text{share, company, business}\}| + |\{\mathrm{Je}\}| + |REF_{\mathrm{Je}}|}\\
&\quad + \frac{2 \times w_{\text{company}} \times (In(\text{company}, \mathrm{Je}) + In(\text{company}, REF_{\mathrm{Je}}))}{|\{\text{share, company, business}\}| + |\{\mathrm{Je}\}| + |REF_{\mathrm{Je}}|}\\
&\quad + \frac{2 \times w_{\text{business}} \times (In(\text{business}, \mathrm{Je}) + In(\text{business}, REF_{\mathrm{Je}}))}{|\{\text{share, company, business}\}| + |\{\mathrm{Je}\}| + |REF_{\mathrm{Je}}|}\\
&= \frac{2 \times \frac{1}{3} \times (1+1) + 2 \times \frac{1}{3} \times (0+0) + 2 \times \frac{1}{3} \times (0+0)}{3+1+1}\\
&= \frac{1.33}{5} = 0.267,
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Sim}'(\mathrm{D}, \mathbf{Fb}) = \frac{2 \times \frac{1}{3} \times (0+1)}{3+1+1} = 0.133,\\
&\mathrm{Sim}'(\mathrm{D}, \mathbf{Fj}) = 0,\\
&\mathrm{Sim}'(\mathrm{D}, \mathbf{Ka}) = 0.
\end{aligned}$$

Step 6: The LDOCE sense, **interest.1.n.3** is linked to the LLOCE sense, **interest-Je**.

Example 2
Linking an LDOCE sense **interest.1.n.3** to its relevant *Roget's* sense.

Step 1–3: The first three steps are independent of the thesaurus used, therefore the same results as in Example 1 should be obtained.

Step 4: Using *Roget's* categories as word classes in *LinkSense,* we have:

$$CLASS_{\text{interest}} = \{\mathbf{Ab}\ (\textit{Dimensions}),\ \mathbf{Cb}\ (\textit{Inorganic matter}),\ \mathbf{Eb}\ (\textit{Prospective volition}),\ \mathbf{Ei}\ (\textit{Possessive relations})\},$$

The keywords *share, company,* and *business* appear in many *Roget's* sections, but only the following sections are relevant to $CLASS_{\text{interest}}$: **Ei** (*share* and *business*), **Eb** (*business*)

Therefore we have:

$$w_{\text{share/n}} = 1/4,\ w_{\text{company/n}} = 1/5,\ w_{\text{business/n}} = 1/7.$$

Step 5: For simplicity, we ignore the cross-reference information in *Roget's* and base our similarity calculation solely on the *CLASS* information. Thus, we have:

$$\text{Sim}(\text{D}, \mathbf{Bb}) = 0,$$

$$\text{Sim}(\text{D}, \mathbf{Cb}) = 0,$$

$$\text{Sim}(\text{D}, \mathbf{Eb}) = \frac{2 \times \frac{1}{7}}{3+1} = \frac{\frac{2}{7}}{4} = \frac{2}{28} = 0.071,$$

$$\text{Sim}(\text{D}, \mathbf{Ei}) = \frac{2 \times \left(\frac{1}{4} + \frac{1}{7}\right)}{3+1} = \frac{\frac{11}{14}}{4} = \frac{11}{56} = 0.196.$$

Step 6: The LDOCE sense **interest.1.n.3** is linked to *Roget's* sense **interest-Ei.**

3.4 Performance Evaluation of *LinkSense*

An experiment involving the LDOCE and the LLOCE was carried out to assess the effectiveness of the *LinkSense* algorithm (see Table 10). To evaluate the performance of algorithms, we define the ratios of applicability A and precision P as follows:

$$A = \frac{\#(\textit{all labeled definitions})}{\#(\textit{all definitions})}$$

$$P = \frac{\#(\textit{correct labeled definitions})}{\#(\textit{all labeled definitions})}\ .$$

Nearly half of the nominal LDOCE senses for a set of highly polysemous words are linked to their relevant LLOCE sense and topics, with a surprisingly high precision rate of 93%. For the other half, *LinkSense* does not find sufficiently high similarity to warrant a link. That is due primarily (approximately two-thirds) to sense gaps in the LLOCE, rather than inconsistency among the LDOCE definitions.

4. Topical Clustering of MRD Senses as Information Retrieval

In this section, we will describe *TopSense,* an algorithm for clustering dictionary senses. *TopSense* clusters closely related senses by applying IR techniques on the results of running *LinkSense* on an MRD. After *LinkSense* links a substantial portion of MRD senses to thesaurus sense classes, we put all definitions of the senses linked to a particular class together in a document. With such a document of collective definitions, topical clustering of all MRD senses bears a striking resemblance to the IR task of

Table 10
Performance of *LinkSense* algorithm.

Headword	# of Definitions in LDOCE	Linking to the LLOCE				
		Correct	Incorrect	Unknown	Applicability	Precision
bass	5	2	1	2	67%	100%
bow	5	4	0	1	80%	100%
cone	3	3	0	0	100%	100%
country	5	5	0	0	100%	100%
crane	2	2	0	0	100%	100%
duty	2	1	0	1	50%	100%
galley	4	2	0	2	50%	100%
interest	6	4	0	2	67%	100%
issue	8	2	1	5	38%	50%
mole	3	2	0	1	67%	100%
plant	6	1	0	5	17%	100%
position	10	7	0	3	70%	100%
sentence	2	2	0	0	100%	100%
slug	5	2	0	3	40%	100%
space	8	2	0	6	25%	100%
star	9	3	2	4	56%	60%
suit	6	3	0	3	50%	100%
table	7	2	0	5	29%	100%
tank	3	1	0	2	33%	100%
taste	6	1	0	5	17%	100%
total	105	51	4	50	52%	93%

retrieving relevant documents for a given query. We observe that the defining words of a sense S frequently recur in documents relevant to *S*. For instance, consider the following LDOCE sense:

star.1.n.5 a piece of *metal* in this shape for *wear*ing as a mark of office, rank, honour, etc.

We observe that most entries in the LDOCE are like **star.1.n.5**, in that they contain defining words that are also recurring terms in a relevant document. For instance, headwords such as *apron*, *bracelet*, *necklace*, and *tie* are defined by using terms in a document corresponding the LLOCE word class for the topic **Dg** (*Clothes and personal belongings*). Table 11 shows the **Dg** class, including such senses as *apron*, *bracelet*, *necklace*, and *tie*, which are indeed relevant to **star.1.n.5**.

4.1 The Clustered Senses as Documents

With topical clustering of MRD senses cast as an IR task, a wealth of well-understood IR techniques can be utilized, including stopword removal, case folding, stemming, term weighting, and document ranking (Witten, Moffat, and Bell 1994). Using the IR analogy, topical clustering of an MRD sense *S* is finding relevant documents (topical clusters), given a query (*S*, the sense definition). With this in mind, we treat the collective definitions of each topical cluster as a virtual document (VD) and reduce the clustering task to ranking relevancy based on terms in the sense definition as well as those in the VDs. For simplicity, we adopt a common scheme of *tf* × *idf* to weight terms in the documents. Each defining term in a VD is associated with a term frequency (*tf*) and document frequency (*df*). Let tf_{ij} represent the frequency of term t_j in document VD_i, and df_j represent the number of VDs where term t_j appears. The relevancy of VD_i to the sense S according to term t_j is therefore given by the following weight:

$$W_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log(N/df_j),$$

where N is the number of documents in the collection. The relevancy of a VD_i to a query Q is obtained by summing up the weights of all terms t_j in Q:

$$\sum_{t_j} tf_{ij} \times \log(N/df_j).$$

Table 11 shows the LDOCE senses and their definitions that are linked to relevant LLOCE senses under a certain topic. An implementation of *LinkSense* links 455 LDOCE senses including *accessory, bracelet,* and *tie* to a **Dg** class in the LLOCE. The definitions of these senses in the **Dg** class form the virtual document $D_{\mathbf{Dg}}$. As shown in Table 12, significantly topical terms are used within a VD consistently. For instance, the term *angle* appears consistently in 10 senses in $D_{\mathbf{Jb}}$ with a weight of 23.82. Table 12 displays heavily weighted terms in some VDs and their associated values of *tf*, *df*, and weight.

4.2 The *TopSense* Algorithm

We sum up the above descriptions and outline the *TopSense* algorithm here.

Algorithm *TopSense*: Topical clustering of MRD senses.

Step 1: Run *LinkSense* on the MRD and thesaurus and collect terms to form VDs.

Step 2: Read sense definition S from the MRD.

Step 3: Remove all stopwords in S and produce a list, Q, of stemmed keywords with part of speech.

Step 4: For each term t_j in Q, look up the corresponding W_{ij} for all virtual documents D_C, $C \in CLASS$, the set of all word classes in the thesaurus.

Step 5: For $C \in CLASS$, calculate $\mathrm{Sim}(Q, D_C) = \sum_{t_j} W_{ij}$, where $W_{ij} = tf_{ij} \times \log(N/df_j)$.

Step 6: Assign S to the class C such that $\mathrm{Sim}(Q, D_C)$ is the largest for all $C \in CLASS$ and passes a preset threshold θ.

4.3 Illustrative Examples: Topical Clustering of LDOCE Senses

Two examples are given in this subsection to illustrate how *TopSense* works. Example 3 shows a calculation done in *TopSense* to find the most relevant topics for another *star* sense (*a 5-or more pointed figure*). Example 4 shows the same calculation done for the sense of *star* (*a piece of metal in this shape for wearing as a mark of office, rank, honour etc.*) discussed above at the beginning of Section 4. Despite very ambiguous terms such as *to wear* (*to dress* or *to rub*) and *figure* (*body, shape,* or *number*) present in both definitions, the weighting scheme of *TopSense* seems to work well enough to determine the relevant topics, **Dg** (*Clothes and personal belongings*) and **Jb** (*Mathematics*), respectively.

Table 11
Partial list of LDOCE senses linked to LLOCE classes by *LinkSense*.

Cluster/ Size	Headword	Sense Definition
Dg / 455	• accessory	• something which is not a necessary part of something larger but which makes it more beautiful, useful, effective etc.
(*Clothes and personal belongings*)	• apron	• a simple garment **worn** over the front part of one's clothes to keep them clean while working or doing something dirty or esp. while cooking.
	• bracelet	• a band or ring, usu. of **metal**, **worn** round the wrist or arm as an ornament.
	• coat	• an outer garment with long SLEEVEs, often fastened at the front with buttons, and usu. **worn** to keep warm or for protection.
	• necklace	• a string of jewels, BEADs, PEARLs, etc., or a chain of gold, silver, etc., **worn** around the neck as an ornament esp. by women.
	• tie	• a band of cloth **worn** round the neck, usu. inside a shirt collar and tied in a knot at the front.
Jb / 212	• cross	• a figure or mark formed by one straight line crossing another, as X, often used.
(*Math*)	• diameter	• a straight line going from side to side through the centre of a circle or other curved figure.
	• pyramid	• a solid figure with a flat usu. square base and straight flat 3-angled sides that slope upwards to meet at a point.
	• rectangle	• a figure with 4 straight sides forming 4 right **angles**.
	• square	• a figure with 4 equal sides and 4 right **angles**.
	• triangle	• a flat figure with 3 straight sides and 3 **angles**.
Ld / 524	• bank	• land along the side of a river, lake, etc.
(*Geography*)	• bay	• a wide opening along a coast; part of the sea or of a large lake enclosed in a curve of the land.
	• beach	• a shore of an ocean, sea, or lake or the bank of a river covered by sand, smooth stones, or larger pieces of rock.
	• lake	• a large mass of water surrounded by land.
	• cascade	• a steep high usu. small waterfall, esp. one part of a bigger waterfall.
Je / 181	• account	• a record or statement of money received and paid out, as by bank or business, esp. for a particular period or at a particular date.
(*Banking*)	• asset	• something such as a house or furniture, that has value and that may be sold to pay a debt.
	• bank	• a place in which money is kept and paid out on demand, and where related activities go on.
	• capital	• wealth, esp. when used to produce more wealth.
	• stock	• money lent to a government at a fixed rate of interest.

Example 3
Clustering an LDOCE sense **star.1.n.3**.

Step 1: Refer to Table 11 for some of the results of running *LinkSense*.

Step 2: S = *"a 5-or more pointed figure."*

Step 3: Q = {pointed/a, figure/n}

Table 12
Some examples of virtual documents, terms, *tf*, *df*, and weight.

VD	Terms	*tf*	*df*	Weight	VD	Terms	*tf*	*df*	Weight
Dg	garment	43	12	102.45	**Ld**	sea	38	23	65.81
	wear	60	27	94.30		land	47	40	55.39
	dress	21	5	68.42		mountain	16	7	46.74
	woman	49	36	62.91		water	44	47	44.76
	coat	19	7	55.51		river	17	15	36.71
	trouser	11	2	45.91		tide	7	1	34.07
	shirt	12	3	45.22		valley	8	2	33.39
	undergarment	8	2	33.39		ocean	9	4	31.33
	shoe	12	11	29.63		lake	10	8	27.88
	skirt	6	1	29.20		shore	8	4	27.84
	cloth	16	25	26.37		earth	16	26	25.75
	waist	8	5	26.06		island	6	2	25.04
	jacket	6	2	25.04		rock	11	14	24.51
	sleeve	5	1	24.33		wave	9	13	20.72
	glove	5	2	20.87		hill	8	10	20.51
	sock	5	2	20.87		deep	12	25	19.78
	neck	10	17	20.34		coast	6	5	19.54
	underpants	4	1	19.47		slope	7	8	19.51
	woolen	5	4	17.40		cliff	5	3	18.84
	tie	7	14	15.59		map	7	9	18.69
Jb	mathematics	15	4	52.21	**Je**	money	42	39	50.56
	multiply	8	3	30.15		account	16	9	42.72
	figure	13	19	25.00		bank	16	12	38.12
	straight	12	17	24.41		pay	17	31	24.37
	angle	10	12	23.82		lend	7	5	22.80
	line	21	44	22.75		interest	12	25	19.78
	circle	9	11	22.22		debt	5	3	18.84
	geometry	5	3	18.84		sum	6	10	15.38
	calculate	7	9	18.69		wealth	5	6	15.37
	add	8	13	18.42		credit	3	1	14.60
	subtract	3	1	14.60		property	5	9	13.35
	curved	7	5	11.54		deposit	2	1	9.73
	perpendicular	2	1	9.73		savings	2	1	9.73
	proportion	2	1	9.73		payment	4	14	8.91
	right-angled	2	1	9.73		share	4	15	8.63
	triangle	2	1	9.73		record	4	16	8.37
	edge	6	26	9.65		spend	3	11	7.40
	arc	2	2	8.34		business	6	40	7.07
	curve	3	9	8.01		supply	4	25	6.59
	cross	3	11	7.40		amount	6	46	6.23

Step 4: For each term in *Q*, we have:

$$W_{\text{pointed},\mathbf{Jb}} = 0, \quad W_{\text{pointed},\mathbf{Kf}} = 0, \quad W_{\text{pointed},\mathbf{Gd}} = 0, \quad W_{\text{pointed},\mathbf{Hd}} = 0,$$
$$W_{\text{figure},\mathbf{Jb}} = 25.00, \quad W_{\text{figure},\mathbf{Kf}} = 9.62, \quad W_{\text{figure},\mathbf{Gd}} = 7.69, \quad W_{\text{figure},\mathbf{Hd}} = 5.77.$$

Step 5: Adding up the weights for each VD, we get

$$\text{Sim}(Q, \mathbf{Jb}) = 25.00,$$
$$\text{Sim}(Q, \mathbf{Kf}) = 9.62,$$
$$\text{Sim}(Q, \mathbf{Gd}) = 7.69,$$

$$\mathrm{Sim}(Q, \mathbf{Hd}) = 5.77.$$

Step 6: For the most relevant topics to S, we get the following ranked list

$$\begin{array}{l}
\mathbf{Jb}\ (\textit{Mathematics}), \\
\mathbf{Kf}\ (\textit{Indoor games}), \\
\mathbf{Gd}\ (\textit{Communicating}), \\
\mathbf{Hd}\ (\textit{Equipment, machines, and instruments}).
\end{array}$$

Example 4
Clustering an LDOCE sense **star.1.n.4.**

Step 1: See Table 11.

Step 2: S = *"a piece of metal in this shape for wearing as a mark of office, rank, honour etc."*

Step 3: Q = {metal/n, shape/n, wear/v, mark/n, office/n, rank/n, honour/n}

Step 4: For each term in Q, we have:

$$\begin{array}{llll}
W_{\text{metal},\mathbf{Dg}} = 3.77, & W_{\text{metal},\mathbf{Hc}} = 62.83, & W_{\text{metal},\mathbf{Ci}} = 0, & W_{\text{metal},\mathbf{Hb}} = 22.62, \\
W_{\text{shapel},\mathbf{Dg}} = 5.98, & W_{\text{spape},\mathbf{Hc}} = 8.97, & W_{\text{shape},\mathbf{Ci}} = 0, & W_{\text{shape},\mathbf{Hb}} = 7.97, \\
W_{\text{wear},\mathbf{Dg}} = 94.30, & W_{\text{wear},\mathbf{Hc}} = 0, & W_{\text{wear},\mathbf{Ci}} = 0, & W_{\text{wear},\mathbf{Hb}} = 4.72, \\
W_{\text{mark},\mathbf{Dg}} = 1.11, & W_{\text{mark},\mathbf{Hc}} = 3.32, & W_{\text{mark},\mathbf{Ci}} = 0, & W_{\text{mark},\mathbf{Hb}} = 6.64, \\
W_{\text{office},\mathbf{Dg}} = 0, & W_{\text{office},\mathbf{Hc}} = 1.61, & W_{\text{office},\mathbf{Ci}} = 3.22, & W_{\text{office},\mathbf{Hb}} = 1.61, \\
W_{\text{rank},\mathbf{Dg}} = 1.69, & W_{\text{rank},\mathbf{Hc}} = 0, & W_{\text{rank},\mathbf{Ci}} = 72.65, & W_{\text{rank},\mathbf{Hb}} = 0, \\
W_{\text{honour},\mathbf{Dg}} = 0, & W_{\text{honour},\mathbf{Hc}} = 0, & W_{\text{honour},\mathbf{Ci}} = 2.30, & W_{\text{honour},\mathbf{Hb}} = 2.30.
\end{array}$$

Step 5: Adding up the weights for each VD, we get

$$\begin{array}{l}
\mathrm{Sim}(Q, \mathbf{Dg}) = 115.16, \\
\mathrm{Sim}(Q, \mathbf{Hc}) = 100.29, \\
\mathrm{Sim}(Q, \mathbf{Hb}) = 88.83, \\
\mathrm{Sim}(Q, \mathbf{Ci}) = 78.17.
\end{array}$$

Step 6: For the most relevant topics to S, we get the following ranked list:

$$\begin{array}{l}
\mathbf{Dg}\ (\textit{Clothes and personal belongings}), \\
\mathbf{Hc}\ (\textit{Specific substances and materials}), \\
\mathbf{Hb}\ (\textit{Object generally}), \\
\mathbf{Ci}\ (\textit{Social classifications and situations}).
\end{array}$$

4.4 Experimental Results

An experiment was conducted to assess the effectiveness of the *LinkSense* and *TopSense* algorithms. The experimental results show that the *LinkSense* links nearly 11,045 of some 39,000 nominal LDOCE senses to a topical sense in the LLOCE. Evaluation based on a 20-word test set shows that, on the average, 50% of the LDOCE instances linked to an LLOCE sense, and, of these links, 95% are correct. These linked LDOCE senses establish 129 topical clusters, one for each LLOCE topic. When the proposed

LinkSense algorithm is applied to assign sense definitions in LDOCE with relevant topical labels, it obtains very high precision but low coverage. *TopSense* is design specifically to improve coverage by providing a reliable method for clustering MRD entries left unlabeled by *LinkSense*.[3] A document of defining terms is then formed from MRD senses in each of these clusters. Subsequently, *TopSense* runs on the nominal LDOCE sense, attempting to merge it to one of the topical clusters.

The thresholds for *LinkSense* and *TopSense* are selected according to random sampling from definitions in the LDOCE. Assume θ is the threshold and $\hat{\theta}$ is an estimator of θ, and B is the bound on the error of estimation. The problem is to limit the error of estimation below B with probability $1-\alpha$. This can be stated as $P(|\theta-\hat{\theta}| < B) = 1-\alpha$, since the number of definitions is large enough to permit estimation of population parameter θ. Considering Central Limit Theory, the parameter $\hat{\theta}$ tends to have approximately a normal distribution. We will usually select $B = 2\sigma_{\hat{\theta}}$, and hence $1 - \alpha$ will be approximately 0.95 for normal distribution. To estimate $\hat{\theta}$, a simple random sample of 100 definitions (about 350 senses) is used. Thus, the estimate of threshold is 0.12 for *LinkSense*. Similar estimation was done for the threshold used in *TopSense*.

Evaluation was done on a set of 20 polysemous words that have been used in recent literature on WSD. These words focus on the more difficult cases of sense ambiguity, as can be seen by the degree of ambiguity as recorded in the LDOCE. These words have 5.3 senses on the average, as opposed to the average of 2.6 senses for all words in the LDOCE.

The evaluation is based on the relevancy assessment by two human judges. The Appendix gives a sense-by-sense rundown of all senses tested and evaluated. Table 13 summarizes the word-by-word applicability and precision of *TopSense*. Although not all senses are clustered and not all clustered senses are correct, applicability and precision are rather high, which seems to indicate that the resulting sense division is directly usable in WSD, and thus, eliminates the need for human intervention.

5. Discussion

In this section, we thoroughly analyze the experimental results, in particular, the cases for which *TopSense* fails. These cases reveal the strengths and limitations of *TopSense* and hint at possible improvements to the algorithm. In addition, we also point out several uses of the topical clusters.

5.1 Failure of the *TopSense* Algorithm

Failure of *TopSense* can be attributed to a number of factors, including vagueness of definitions, inappropriate definition lengths (too short or too long), metaphoric or metonymic senses, and deictic references. Table 14 shows some examples of the failed cases. For instance, the sense **interest.1.n.3** (*a readiness to give attention*) is too vague and short for correct clustering to occur. On the other hand, long definitions including too many non-essential differentiae also give rise to erroneous clustering. We notice that the definitions of such senses have been radically changed and made more specific in the third edition of the LDOCE. The reason behind the changes may be that these sense definitions are also difficult for humans to grasp.

Metonymic senses sometimes lead to problems for the proposed algorithms. *TopSense* successfully puts **start.1.n.3** (*a piece of metal in this shape for wearing as a mark*

[3] It seems that precision may be lower if *TopSense* is run on the unlabeled entries, but we suspect the difference is very small.

Table 13
Evaluation of the *TopSense* algorithm.

Headword	#of Definitions	Labeling with Expanded Candidate Set				
	in LDOCE	Correct	Incorrect	Unknown	Applicability	Precision
bass	5	5	0	0	100%	100%
bow	5	5	0	0	100%	100%
cone	3	3	0	0	100%	100%
country	5	5	0	0	100%	100%
crane	2	2	0	0	100%	100%
duty	2	2	0	0	100%	100%
galley	4	4	0	0	100%	100%
interest	6	4	2	0	100%	67%
issue	8	3	2	3	63%	60%
mole	3	3	0	0	100%	100%
plant	6	5	1	0	100%	83%
position	10	9	1	0	100%	90%
sentence	2	2	0	0	100%	100%
slug	5	5	0	0	100%	100%
space	8	7	1	0	100%	88%
star	9	8	1	0	100%	90%
suit	6	5	1	0	100%	83%
table	7	6	1	0	100%	86%
tank	3	3	0	0	100%	100%
taste	6	5	0	1	83%	100%
total	105	91	10	4	96%	90%

of office, rank, honour, etc.) in the **Dg** class (*Clothes and personal belongings*). On the other hand, the metonymic meaning, **Nb** (*Chance*) of another *star* sense (*a heavenly body regarded as determining one's fate*) comes out second to the "primary" sense, **La** (*Heavenly body*). By considering cue phrases such as *regarded as* or *as a mark of*, we might be able to handle metaphoric and metonymic senses more successfully.

Krovetz (1992) observes that the LDOCE indicates explicit sense shifts via the deictic reference, which is a link to the previous sense created by such terms as *this, these, that, those, its, itself, such a,* and *such an*. The author identifies many systematic sense shifts indicated by such references including **Substance/Product** (*lemon, tree or fruit*), **Substance/Color** (*jade, amber*), **Object/Shape** (*pyramid*), **Animal/Food** (*chicken*), **Count-noun/Mass-noun** (*blasphemy*), **Language/People** (*Spanish*), **Animal/Skin-fur** (*crocodile*), and **Music/Dance** (*waltz*). Such shifts indicated through a deictic reference are so pervasive in the MRD that they show up more than once in our small 20-word test set. For instance, the LDOCE sense **issue.1.n.2** (*an example of this*) indicates a **Count-noun/Mass-noun** shift from its previous sense **issue.1.n.1** (*the act of coming out*) through the deictic reference of *this*. Since these specific patterns of definition are not taken into consideration in *TopSense*, the algorithm often fails in such cases. Further work must be undertaken to cope with direct and deictic references, so that such definitions can be appropriately clustered.

5.2 Clustered Definitions and Examples as a Knowledge Source for WSD

Many studies have shown that MRD definitions and example sentences are a good knowledge source for WSD. As described in the introduction, Lesk (1986) shows that defining words are especially effective for disambiguating senses strongly associated

Table 14
Analysis of failure by error types.

Error Type	*TopSense* Output	Sense Definition
vague definition	*Gd (communicating)	**interest** - an activity, subject, etc., which one gives time and attention to
long definition	*Ca (people)	**table** - also multiplication table; a list which young children repeat to learn what number results when a number from 1 to 12 is multiplied by any of the numbers from 1 to 12
metonynym	* La (universe)	**star** - a heavenly body regarded as determining one's fate
short, vague definition	*Ge (communication)	**suit** - a set (of armour)
	*Bj (medicine)	**interest** - a readiness to give attention
	– (unknown)	**issue** - the act of coming out
	– (unknown)	**issue** - something which comes or is given out
deictic reference	*Hb (object)	**space** - a quantity or bit of this for a particular purpose
	– (unknown)	**issue** - an example of this

with specific collocations, such as *cone* in *ice-cream cone* and *pine cone.* Wilks et al. (1990) call the defining words in the LDOCE definition **semantic primitives** (SP) and suggest that a semantic network constructed on the strength of co-occurrence of SPs in definitions can be useful for a variety of NLP tasks, ranging from WSD, to machine translation, to message understanding. Along the same lines, Luk (1995) terms SP the **definition-based concept** (DBC) and proposes using DBC co-occurrence (DBCC) trained on a large corpus to disambiguate word senses. However, the effectiveness of SPs or DBCs to represent a word sense and its indicative context is hampered by ambiguity and data sparseness. For instance, *earth,* one of the SPs in **bank.1.n.2** is ambiguous (either as *the planet Earth* or *soil*) thus possibly leading to problems in WSD. Although these SPs are drawn from a small, controlled vocabulary in most MRDs, nevertheless, it is difficult to find SPs of a polysemous sense overlapping the SPs of its context. For instance, consider the problem of disambiguating the word *bank* in the context of an LDOCE example, *He sat down and rested on a mossy bank in the woods.*

When working on the level of the SPs of an individual MRD sense, we are hard pressed to find a match between the SPs of the intended sense:

$$\mathrm{SP}(\mathbf{bank.1.n.2}) = \{\mathrm{earth, heap, field, garden, make, border, division}\}$$

and the SPs of its context:

$$\begin{aligned}
\mathrm{SP}(\mathbf{sit}) &= \{rest, position, upper, body, upright, support, bottom, chair, seat\},\\
\mathrm{SP}(\mathbf{rest}) &= \{take, rest\},\\
\mathrm{SP}(\mathbf{moss}) &= \{small, flat, green, yellow, flowerless, plants, grow, thick, furry, wet,\\
&\quad \mathrm{soil, surface}\},\\
\mathrm{SP}(\mathbf{wood.1.n.1}) &= \{material, trunk, branch, tree, cut, dry, form, burn, paper, furniture\},\\
\mathrm{SP}(\mathbf{wood.1.n.2}) &= \{place, tree, grow, small, forest\}.
\end{aligned}$$

The clusters of MRD senses produced by *TopSense* give us an advantage in this respect. By matching the context against the clustered semantic primitives (CSP) of the related senses, we have a better chance of a match. For instance, the following CSPs of the relevant *bank* senses contains more words, therefore are more likely to recur in the SPs of contextual words:

$$\begin{aligned}\text{CSP}(\textbf{bank-Ld}) &= \text{SP}(\textbf{bank.1.n.1}) \cup \text{SP}(\textbf{bank.1.n.2}) \cup \text{SP}(\textbf{bank.1.n.4}) \\ &\quad \cup \text{SP}(\textbf{bank.1.n.5}) \\ &= \{\text{hand, side, river, stream, lake, earth, heap, field, make, border,} \\ &\quad \text{division, slope, bend, road, race-track, safe, car, go round,} \\ &\quad \text{sandbank}\}\end{aligned}$$

If data sparseness still gets in the way, as in the case of this example, one can go one step further and adopt a class-based approach. Under such an approach, the SPs of the context are matched against the SPs of a class of senses related to the polysemous sense in question. To this end, we can make use of the topical clusters of MRD senses produced by *TopSense*. By taking the collective defining terms of all the senses in a topical cluster, we obtain the virtual document of SPs described in Section 4.1. To cope with the problem caused by ambiguous SPs, it is a good idea to weight terms according to *tf* and *idf*, as in the *TopSense* algorithm. Under such a class-based approach, we will be matching the contextual information against the unweighted or weighted terms in a class relevant to the intended sense. For instance, to resolve the sense of *bank* in the above example to the **Ld** sense, we look for a match of contextual information with V_{Ld}.

$$\begin{aligned}\text{V}_{\text{Ld}} &= \text{CSP}(\textbf{bank-Ld}) \cup \text{CSP}(\textbf{forest-Ld}) \cup \text{CSP}(\textbf{valley-Ld}) \cup \cdots \\ &= \{\text{land, side, river, stream, lake, earth, heap, field, make, border, division,} \\ &\quad \text{slope, bend, road, race-track, safe, car, go round, sandbank, large,} \\ &\quad \text{area, land, thick, cover, } \textit{tree}\text{, bush, grow, wild, plant,} \\ &\quad \text{purpose,} \ldots\} \\ \text{V}_{\text{Ld}} &= \{\text{sea } (65.81), \text{land } (55.39), \text{mountain } (46.74), \text{water } (44.76), \text{river } (36.71), \\ &\quad \text{lake } (27.88), \text{earth } (25.76), \textit{tree } (21.87), \ldots\}\end{aligned}$$

Notice that for this example, the relevant VD is now large enough to overlap the contextual information; the term *tree* appears in SP (**wood.1.n.1**) as well as the relevant document V_{Ld}. Although the relevant V_{Ld} is very large, it contains mostly words that are nevertheless consistently related to geography.

5.3 Systematic Sense Shift

Ostler and Atkins (1991) contend that there is strong evidence to suggest that a large part of word sense ambiguity is not arbitrary but follows regular patterns. Moreover, gaps frequently arise in dictionaries and thesauri in specifying this kind of polysemy. Encoding regularity of the extended usage of a sense makes it possible to resolve word sense ambiguity for word entries that are underspecified in this respect. This so-called virtual polysemy can be illustrated through some examples. For instance, many verbs for moving and action, such as *move* and *strike*, can be used polysemously in the sense of emotion. Chodorow, Byrd, and Heidom (1985) observe that many instances of intersense relations can be found in W7 that are not idiosyncratic, but rather exist among senses of many words. Those relations include **Process/Result**, **Food/Plant**, and **Container/Volume**. Virtual polysemy and recurring intersense relations are closely related to polymorphic senses that can support coercion in semantic typing under Putstejovsky's (1991) theory of the generative lexicon.

Dolan (1994) maintains the position that intersense relations are mostly idiosyncratical, thereby making it difficult to characterize them in a general way so as to identify them. The author cites the example of two senses of *to moult*, one a bird behavior and the other an animal behavior, to stress that polysemy primarily reflects fine distinctions that do not recur systematically throughout the English lexicon. However, our experimental results indicate that (a) it is exactly senses with fine distinction that are merged together and (b) there is a greater concentration of recurring intersense relations emerging from condensed senses. For instance, the distinction between the bird and animal behavior of *moulting* would be eliminated, since both are likely to be clustered and labeled as **Ha** (*Making things*) by *TopSense*. Relations among senses in the same topical clusters are mostly systematic. Many of those relations are reflected in the cross-reference information in the LLOCE. For instance, the LLOCE lists the following cross-references for the topic of **Eb** (*Food*):

Ac (*Animals/Mammals*),

Ad (*Birds*),

Af (*Fish and other water creatures*),

Ah (*Parts of animal*),

Ai (*Kinds of parts of plants*),

Aj (*Plant in general*),

Jg (*Shopkeepers and shops selling food*).

Most of those cross-references are systematic intersense relations similar to the abovementioned **Food/Plant** relation. Indeed, words involved in such intersense relations are frequently underspecified. For instance, *chicken* is listed under both topic **Eb** and topic **Ad**, while *duck* is listed under **Ad** but not **Eb**.

By characterizing some 200 cross-references in LLOCE, most systematic sense shifts can be easily identified among the senses across topical clusters. The topical clusters of MRD senses, coupled with the topical description of sense-shift knowledge, can support and realize automatic sense extension, as advocated in Putstejovsky and Bouillon (1994), and prevent a proliferation of senses in the semantic lexicon. For instance, the sense of *duck* in the **Ad** cluster can be coerced into an **Eb** sense, in some context, based on the knowledge of a systematic sense shift from **Ad** (*Birds*) to **Eb** (*Food*).

6. Other Approaches

Sanfilippo and Poznanski (1992) propose a so-called dictionary correlation kit (DCK) in a dialogue-based environment for correlating word senses across a pair of MRDs such as the LDOCE and the LLOCE. The approach taken in DCK is essentially a heuristic one, based on a correlation in the headwords, grammar codes, definition, and examples between the senses in LDOCE and LLOCE. The authors indicate that for the heuristics to yield optimum results, the degree of overlap in the examples should be weighted twice as heavily as all other factors. However, they do not elaborate on how the comparisons are done, or on how effective the program is.

Dolan (1994) describes a heuristic approach to forming unlabeled clusters of closely related senses in an MRD. The clustering program relies on LDOCE domain code, grammar code, and 25 types of semantic relations extracted from definitions such as *Hypernym, Location, Manner, Purpose, PartOf,* and *IngredientOf.* Matching two senses involves comparing any values that have been identified for each of the semantic relation types. The author reports that straightforwardly comparing the values of the same semantic relation types, particularly the *Hypernym* relation, for two senses would be quite effective. In addition to such a comparison, a number of "scrambled" comparisons between values of different types of semantic relations are also helpful. For instance, in comparing the two senses of *coffee*, the value "drink" in the sense, "the coffee as a drink" is compared with that of the *IngredientOf* relation in another sense, "the powder as an ingredient of the drink."

Yarowsky (1992) describes a WSD method and an implementation based on *Roget's Thesaurus* and a very large corpus, the 10-million-word *Grolier's Encyclopedia*. He suggests that the method can be applied to disambiguation and merging of MRD definitions as well, and gives the results of applying the method to the senses of the word *crane* for the COBUILD and *Collins* dictionaries using *Roget's* categories as an example. It is not known how the method fares for words other than *crane*. Contrary to our approach, the method requires substantial data for training.

In most of the above-mentioned works, experimental results are reported only for some senses of a few words. In this study, we have evaluated our method using all senses for 20 words that have been studied in WSD literature. This evaluation provides an overall picture of the expected success rate of the method when applied to all word senses in the MRD. Direct comparison of methods is often difficult, but it is clear that, as compared to other methods discussed above, our algorithm is very simple, requires minimal preprocessing, and does not rely on information idiosyncratic to the MRD, such as the LDOCE subject code or grammar code. Thus, the algorithm described in this paper can be readily applied to other MRDs besides LDOCE. Although our algorithm makes use of defining words in various semantic relations with the sense, those relations need not be explicitly computed through an elaborated parsing and extraction process.

Finally, it is interesting to compare our method with some aspects of the program for induction of sense division of Schütze (1992). As mentioned in the introduction, the program uses distributional similarity of lexical co-occurrence to partition word instances into clusters that are likely to be related to sense division. Drawing on the work of latent semantic indexing in IR research, words and contexts are represented as vectors in a multidimensional space. Regression techniques of singular value decomposition are used to reduce the representation to a lower dimensional space. After that, sense division is derived through unsupervised clustering of these word instances. Our method, on the other hand, relies primarily on co-occurrence in an existing set of topical clusters, the topics in LLOCE or *Roget's*. The sense in question is simply merged to the nearest topical cluster. Low-cost distance calculation is done according to the overlap between words in a definition and a topical cluster.

7. Conclusions and Future Work

This paper presents the issues of WSD using machine-readable dictionaries. It describes simple but effective algorithms for disambiguating and clustering dictionary senses to create a sense division for WSD. The proposed algorithms are effective for specific linguistic reasons. Although word sense is an abstract concept that relies on the subjective and subtle distinction of many factors, coarse word sense division can be attributed primarily to the subject and topic. This is evident from the observation that very topical genus and differentiae show up in dictionary definitions in rather rigid patterns. Therefore, an MRD coupled with a thesaurus organized according to subjects and topics is very effective for acquisition of sense division for WSD.

In a broader context, this paper presents an approach to automatic construction of semantic lexicons through integration of lexicographic resources such as MRDs and thesauri. As noted in Dolan (1994), it is possible to run a sense-clustering algorithm on several MRDs to build an integrated lexical database with more complete coverage of word senses. If *TopSense* is run on several bilingual MRDs, there is a potential for creating an integrated multilingual lexicon enriched with thesaurus concepts as language-neutral signs to support knowledge-based machine translation. A similar idea has been put forward by Okumura and Hovy (1994).

The *TopSense* algorithm's performance could definitely be improved by handling deictic, metonymic, and metaphoric sense definitions more appropriately. Nevertheless, the algorithm already produces clustered MRD sense entries that not only are exploitable as a workable sense division but also are likely to be an effective knowledge source for many NLP tasks related to semantic processing, such as WSD. In summary, this paper presents a functional core for automatic construction of the semantic lexicon.

Appendix

The following table shows the experimental results of running *TopSense* on the LDOCE senses in a test set of 20 highly polysemous words.

bass

Topical Clustering	Definition Sentences	Applicability	Precision
• Eb (food)	• any of many kinds of fresh-water or salt-water fish that have prickly skins and that can be eaten.	100%	100%
• Gd (communicating)	• the lowest part in written music.		
• Kb (music)	• the lowest male singing voice.		
• Kb (music)	• a deep voice.		
• Kb (music)	• = DOUBLE BASS.		

bow

Topical Clustering	Definition Sentences	Applicability	Precision
• Dg (clothes and personal belongings)	• a knot formed by doubling a line into 2 or more round or curved pieces, and used for ornament in the hair, in tying shoes, etc.	100%	100%
• Hc (substances and materials)	• a piece of wood held in a curve by a tight string and used for shooting arrows.		
• Kb (music)	• a long thin piece of wood with a tight string fastened along it, used for playing musical		
	• instruments that have strings.		
• Ma (moving)	• a bending forward of the upper part of the body to show respect or yielding.		
• Mf (shipping)	• the forward part of a ship.		

cone

Topical Clustering	Definition Sentences	Applicability	Precision
• Aj (plants)	• the fruit of a PINE or FIR, consisting of several partly separate seed-containing pieces laid over each other, shaped rather like this.	100%	100%
• Hb (objects)	• a hollow or solid object shaped like this.		
• Hf (containers)	• a solid object with a round base and a point at the top.		

country

Topical Clustering	Definition Sentences	Applicability	Precision
• Ld (geography)	• a nation or state with its land or population.	100%	100%
• Ce (organization)	• the nation or state of one's birth or citizenship.		
• Ce (organization)	• the people of a nation or state.		
• Ld (geography)	• land with a special nature or character		
• Ld (geography)	• the land outside cities or towns; land used for farming or left unused.		

crane

Topical Clustering	Definition Sentences	Applicability	Precision
• Hd (equipments)	• a machine for lifting and moving heavy objects by means of a very strong rope or wire fastened to a movable arm (JIB).	100%	100%
• Ad (birds)	• a type of large tall bird with very long legs and neck, which spends much time walking in water catching fish in its very long beak.		

duty

Topical Clustering	Definition Sentences	Applicability	Precision
• Jf (commerce)	• any of various types of tax.	100%	100%
• Jh (work)	• what one must do either because of one's job or because one thinks it right		

galley

Topical Clustering	Definition Sentences	Applicability	Precision
• Gd (communicating)	• a long flat container used by a printer to hold the letters (TYPE) which have been arranged for the first stage of printing.	100%	100%
• Gd (communicating)	• =GALLEY PROOF.		
• Mf (shipping)	• a ship which was rowed along by slaves.		
• Mf (shipping)	• a ship's kitchen.		

interest

Topical Clustering	Definition Sentences	Applicability	Precision
• *Bj (medicine)	• a readiness to give attention.	100%	67%
• *Gd (communicating)	• an activity, subject, etc., which one gives time and attention to.		
• Je (banking)	• advantage, advancement, or favour (esp. in the phrs. in the interest of (something)/in someone's interest).		
• Je (banking)	• money paid for the use of money.		
• Jf (commerce)	• a share (in a company, business, etc.		
• Na (being, becoming)	• a quality of causing attention to be given.		

issue

Topical Clustering	Definition Sentences	Applicability	Precision
• – (unknown)	• the act of coming out.	63%	60%
• – (unknown)	• something which comes or is given out.		
• – (unknown)	• an important point.		
• Ca (people)	• old use and law children (esp. in the phr. die without issue).		
• *Ck (courts of law)	• the act of bringing out something in a new form.		
• *Cl (police, crime)	• an example of this.		
• Gd (communicating)	• something, esp. something printed, brought out again or in a new form.		
• Nf (causing)	• the result.		

	mole		
Topical Clustering	Definition Sentences	Applicability	Precision
• Ac (animals)	• a small, dark brown, slightly raised mark on a person's skin, usu. there since birth.	100%	100%
• Hc (specific substances and materials)	• a type of small insect-eating animal with very small eyes and soft dark fur, which digs holes and passage underground and makes its home in them.		
• Ld (geography)	• a stone wall of great strength built out into the sea from the land as a defense against the force of the waves, or to act as a road.		

	plant		
Topical Clustering	Definition Sentences	Applicability	Precision
• Ai (plants)	• a living thing that has leaves and roots, and grows usu. in earth, esp. the kind smaller than trees.	100%	83%
• Hd (equipment)	• a machine; apparatus.		
• Id (industry)	• a factory ().		
• *Md (vehicles)	• machinery.		
• Ce (organization in groups)	• a person who is placed in a group of people thought to be criminals in order to discover facts about them.		
• Cl (crime)	• a thing, esp. stolen goods, hidden on a person so that he will seem guilty.		

position			
Topical Clustering	Definition Sentences	Applicability	Precision
• Me (places)	• the place where someone or something is or stands, esp. in relation to other objects, places, etc.	100%	90%
• Me (places)	• the place where someone or something is (in the phr. in position).		
• Me (places)	• the place where someone or something is supposed to be; the proper place.		
• Cn (fighting)	• the place of advantage in a struggle (in the phrs. manoeuvre/ jockey for position).		
• Ma (moving)	• the way or manner in which someone or something is placed or moves, stands, sits, etc.		
• *Ca (people)	• a condition or state, esp. in relation to that of someone or something else.		
• Ci (classifications)	• a particular place or rank in a group.		
• Cf (government)	• high rank in society, government, or business.		
• Jh (work)	• a job; employment.		
• Ga (thinking)	• an opinion or judgment on a matter.		

sentence			
Topical Clustering	Definition Sentences	Applicability	Precision
• Ck (courts of law)	• a punishment for a criminal found guilty in court.	100%	100%
• Gd (communicating)	• a group of words that forms a statement, command, EXCLAMATION, or question, usu. contains a subject and a verb, and (in writing) begins with a capital letter and ends with one of the marks ".!?"		

slug

Topical Clustering	Definition Sentences	Applicability	Precision
• Ab (living creatures)	• any of several types of small limbless plant-eating creature, related to the SNAIL but with no shell, that often do damage to gardens.	100%	100%
• Gd (communicating)	• a machine-made piece of metal with a row of letters along the edge for printing.		
• Hc (specific substances)	• a machine-made piece of metal with a row of letters along the edge for printing.		
• Hd (equipment)	• a coin-shaped object unlawfully put into a machine in place of a coin.		
• Hh (weapons)	• a bullet.		

space

Topical Clustering	Definition Sentences	Applicability	Precision
• Jc (measurement)	• something limited and measurable in length, width, or depth and regarded as not filled up; distance, area, or VOLUME (3); room.	100%	88%
• *Hb (objects)	• a quantity or bit of this for a particular purpose.		
• La (universe)	• that which surrounds all objects and continues outward in all directions.		
• La (universe)	• what is outside the earth's air; where other heavenly bodies move.		
• Ld (geography)	• land not built on (esp. in the phr. open space).		
• Le (time)	• a period of time.		
• Gd (communicating)	• an area or distance left between written or printed words, lines etc.		
• Gd (communicating)	• the width of a letter on a TYPEWRITER.		

star

Topical Clustering	Definition Sentences	Applicability	Precision
• Dg (personal belongings)	• a piece of metal in this shape for wearing as a mark of office, rank, honour, etc.	100%	90%
• Jb (mathematics)	• a 5- or more-pointed figure.		
• Kd (drama)	• a famous or very skillful performer.		
• La (universe)	• STARS.		
• La (universe)	• a brightly-burning heavenly body of great size, such as the sun but esp. one very far away.		
• La (universe)	• any heavenly body (such as a PLANET) that appears as a bright point in the sky.		
• *La (universe)	• a heavenly body regarded as determining one's fate.		
• La (universe)	• a sign used with numbers from usu. 1 to 5 in various systems, and in the imagination, to judge quality.		
• Nb (possibility)	• one's success or fame or chance of getting it.		

suit

Topical Clustering	Definition Sentences	Applicability	Precision
• Dg (clothes)	• a set of outer clothes which match, usu. including a short coat (JACKET) with trousers or skirt.	100%	83%
• Dg (clothes)	• a garment or set of garments for a special purpose.		
• *Ge (communication)	• a set (of armour) (in the phrs. suit of armour/mail).		
• Kf (indoor games)	• one of the 4 sets of cards used in games.		
• Gc (communicating)	• fml a request ().		
• Cb (courting)	• old use the act of asking a woman to marry (esp. in the phrs. plead/press one's suit).		

tank

Topical Clustering	Definition Sentences	Applicability	Precision
• Hf (containers)	• a large container for storing liquid or gas.	100%	100%
• Md(vehicles)	• an enclosed heavily armed armoured vehicle that moves on 2 endless metal belts.		
• Ld(geography)	• esp. Ind & PakE a large man-made pool for storing water.		

table

Topical Clustering	Definition Sentences	Applicability	Precision
• Hb (objects)	• a piece of furniture with a flat top supported by one or more upright legs.	100%	86%
• Df (furniture)	• made to be placed and used on such a piece of furniture.		
• Kf (indoor games)	• such a piece of furniture specially made for the playing of various games.		
• Ea (food)	• the food served at a meal.		
• Ca (people)	• the people sitting at a table.		
• Gd (communicating)	• a printed or written collection of figures, facts, or information arranged in orderly rows across and down the page.		
• *Ca (people)	• also multiplication table a list which young children repeat to learn what number results when a number from 1 to 12 is multiplied by any of the numbers from 1 to 12.		

	taste		
Topical Clustering	Definition Sentences	Applicability	Precision
• Bg (bodily states)	• an experience.	83%	100%
• Ea (food generally)	• a small quantity of food or drink.		
• Eb (food)	• the special sense by which a person or animal knows one food from another by its sweet, bitter, salty, etc.		
• Eb (food)	• the sensation that is produced when food or drink is put in the mouth and that makes it different from other foods or drinks by its salty, sweet, bitter, etc.		
• Kb(music)	• the ability to enjoy and judge beauty, style, art, music, etc.; ability to choose and use the best manners, behaviour, fashions, etc.		
• – (unknown)	• a personal liking for something.		

Acknowledgments

This work is partially supported by ROC NSC grants 84-2213-E-007-023 and NSC 85-2213-E-007-042. We are grateful to Betty Teng and Nora Liu from Longman Asia Limited for the permission to use their lexicographical resources for research purposes. Finally, we would like to thank the anonymous reviewers for many constructive and insightful suggestions.

References

Ageno, A., I. Castellon, M. A. Marti, G. Rigau, F. Ribas, H. Rodriguez, M. Taule and F. Verdejo. 1992. SEISD: An environment for extraction of semantic information from on-line dictionaries. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 253–254, Trento, Italy.

Ahlswede, Thomas and Martha Evens. 1988. Parsing vs. text processing in the analysis of dictionary definitions. In *Proceedings of the 26th Annual Meeting*, pages 217–224. Association for Computational Linguistics.

Alshawi, H. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *American Journal of Computational Linguistics*, 13(3):195–202.

Alshawi, H., B. Boguraev, and D. Carter. 1989. Placing the dictionary on-line. In B. Boguraev and Briscoe, editors, *Computational Lexicography for Natural Language Processing*, Longman, London, pages 41–63.

Amsler, Robert A. 1984a. Machine-readable dictionaries. *Annual Review of Information Science and Technology*, 19:161–209.

Amsler, Robert A. 1984b. Lexical knowledge bases, Panel session on machine-readable dictionaries. In *Proceedings of the Tenth International Congress on Computational Linguistics*, pages 458–459, Stanford, CA.

Amsler, Robert A. 1987. Words and words. In *Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing*, pages 7–9, New Mexico State University at Las Cruces, NM.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1991. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting*, pages 264–270. Association for Computational Linguistics.

Bruce, Rebecca and Janyce Wiebe. 1995. Word sense disambiguation using decomposable models. In *Proceedings of the 33rd Annual Meeting*, pages 139–145. Association for Computational Linguistics.

Chang, Jason S., J. N. Chen, H. H. Sheng, and S. J. Ker. 1996. Combining machine readable lexical resources and bilingual corpora for broad word sense disambiguation. In *Proceedings of the Second Conference of the Association for*

Machine Translation, pages 115–124, Montreal, Quebec, Canada.

Chen, J. N. and Jason S. Chang. 1994. Towards generality and modularity in statistical word sense disambiguation. In *Proceedings of the 8th Asian Conference on Language, Information and Computation*, pages 45–48.

Chodorow, Martin S., Roy J. Byrd and George E. Heidom. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting*, pages 299–304. Association for Computational Linguistics.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, Austin, TX.

Copestake, A. 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. In *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*, pages 19–29, Tilburg, The Netherlands.

Cowie, Jim, Joe Guthrie and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 359–365.

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting*, pages 130–137. Association for Computational Linguistics.

Dolan, William B. 1994. Word sense disambiguation: Clustering related senses. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 712–716.

Gale, W., K. W. Church, and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

Guthrie, Louise, Brian M. Slator, Yorick Wilks, and Rebecca Bruce. 1990. Is there contents in empty heads? In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 138–143.

Jenson, Karen and Jean Louis Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics*, 13(3-4):251–260.

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

Krovetz, Robert. 1992. Sense-linking in a machine readable dictionary. In *Proceedings of the 30th Annual Meeting*, pages 330–332. Association for Computational Linguistics.

Krovetz, R. and W. B. Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transaction on Information Systems*, pages 115–141.

Lesk, Michael E. 1986. Automatic sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice-cream cone. In *Proceedings of the ACM SIGDOC Conference*, Toronto, Ontario, pages 24–26.

Proctor, P., editor. 1978. *Longman Dictionary of Contemporary English*. Longman Group, London.

Luk, Alpha K. 1995. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Annual Meeting*, pages 181–188. Association for Computational Linguistics.

McArthur, Tom. 1992. *Longman Lexicon of Contemporary English*. Longman Group (Far East) Ltd., Hong Kong.

McRoy, S. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An on-line lexical database. CSL 43, Cognitive Science Laboratory, Princeton University, Princeton, NJ.

Montemagni, S. and L. Vanderwende. 1992. Structural pattern vs. string pattern for extracting semantic information from dictionaries. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, pages 546–552.

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of the 34th Annual Meeting*, pages 40–47, Santa Cruz, CA. Association for Computational Linguistics.

Okumura, A. and Eduard Hovy. 1994. Lexicon-to-ontology concept association using a bilingual dictionary. In *Proceedings of the First Conference of the Association for Machine Translation in the Americans*, pages 177–184. Columbia, MD.

Ostler, Nicholas and B. T. S. Atkins. 1991. Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Proceedings of the 1991 ACL Workshop on Lexical Semantics and Knowledge Representation*, pages 76–87.

Putstejovsky, James. 1991. The generative lexicon. *Computational Linguistics*, (17)4:409–441.

Putstejovsky, James and Pierrette Bouillon. 1994. On the proper role of coercion in semantic typing. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 706–711.

Ravin, Yael. 1990. Disambiguating and interpreting verb definitions. In *Proceedings of the 28th Annual Meeting*, pages 260–267. Association for Computational Linguistics.

Roget's Thesaurus of English Words and Phrases. 1987. Longman Group UK Limited, London.

Sanfilippo, A. and V. Poznanski. 1992. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92)*, pages 80–87, Trento, Italy.

Schütze, Hinrich. 1992. Word sense disambiguation with sublexical representations. In *Proceedings of the 1992 AAAI Workshop on Statistically-based Natural Language Programming Techniques*, pages 100–104.

Vanderwende, L. 1994. Interpretation of noun sequences. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 454–460.

Vossen, P., W. Meijs, and M. den Broeder. 1989. Meaning and structure in dictionary definitions. In *Computational Lexicography for Natural Language Processing*. Branimir Boguraev and Ted Briscoe, editors, Longman Group UK Limited, London, pages 171–190.

Webster's Seventh New Collegiate Dictionary. 1967. C. and C. Merriam Company, Springfield, MA.

Wilks, Y. A., D. C. Fass, C. Ming Guo, J. E. McDonald, T. Plate, and B. M. Slator. 1990. Providing tractable dictionary tools. *Machine Translation*, 5:99–154.

Witten, Ian H., Alistair Moffat, and Timothy C. Bell. 1994. *Managing Gigabytes*. Van Nostrand Reinhold, New York.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 454–460, Nantes, France.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting*, pages 189–196. Association for Computational Linguistics.

Zernik, Uri. 1992. Train1 vs. Train2: Tagging word senses in corpus. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, pages 91–112.