



Multi-dictionary with word sense disambiguation system architecture

Numtip Rattanawongchaiya, Institute for Innovation and Development of Learning Process, Mahidol University, Thailand,

Kanlaya Naruedomkul, Department of Mathematics, Mahidol University, Thailand and

Nick Cercone, Department of Computer Science, Dalhousie University, Canada

g4637458@student.mahidol.ac.th cknr@mahidol.ac.th nick@cs.dal.ca

Abstract: *The wealth of scientific information published in English presents difficulties to students who learn English as a second language. Indeed even in the study of science, this can be a problem because many scientific terms are not found in dictionaries. Another complication is the selection of meaning from the different senses of words in the dictionary. In this paper, we present a multi-dictionary with word sense disambiguation system architecture. The purpose of the system is to facilitate students in scientific learning in English. The proposed system search from various online dictionaries and choose the appropriate meaning of a word according to a given context. The system architecture is composed of a multi-dictionary and word sense disambiguation modules. In this paper, we present the preliminary experiment of adapting word sense disambiguation system with dictionary. The sample sets are composed of 20 phrases each. The average of accuracy from word sense disambiguation module is 82.5 percent.*

Introduction

There are many challenges for novices in learning English, in particular a new vocabulary, English syntax and the pragmatics of language use. Obstacles increase for students learning science, as many scientific terms are rarely found in dictionaries. Another cause of frustration is selecting the most appropriate meaning of the word from many alternative senses. Searching many dictionaries or encyclopaedias can solve the problem but this takes an unreasonable amount of time for students searching for words manually. A number of existing online dictionaries provide information from various resources (such as onelook.com 2006, answers.com 2006) called multi-dictionary. Nevertheless, existing online multi-dictionaries do not provide meaning selection. The meaning selection task is published in natural language processing (NLP) area called word sense disambiguation (*WSD*) since 1950s (Ide and Veronis 1998). Generally, *WSD* is applied in natural language applications such as information retrieval, machine translation, grammatical analysis, speech processing and so on. In this paper, we propose a multi-dictionary with *WSD* system. With the success of the method proposed by Liu, Yu and Meng (2005), we adapt their *WSD* method to incorporate multi-dictionary. Our aim is to develop a facilitating dictionary that not only collates meanings of a word from various online dictionaries but also selects the appropriate meaning for words according to their context. The experiment outlined in this paper evaluates the accuracy of meaning selection by the *WSD* method.

System architecture

To accomplish our aim, the system is composed of (a) a multi-dictionary module and (b) a word sense disambiguation module as illustrated in Figure 1. The multi-dictionary collates the information from different resources including dictionaries and encyclopaedias, and then presents the information to students. *WSD* integrates with multi-dictionary for meaning selection according to a given context.

The system input is either only a word or a word with its context which is required if meaning selection is requested. When the meaning selection is requested, the input word and its context will be sent to the word sense disambiguation module which uses *WordNet* (Miller 1990) as a knowledge base.

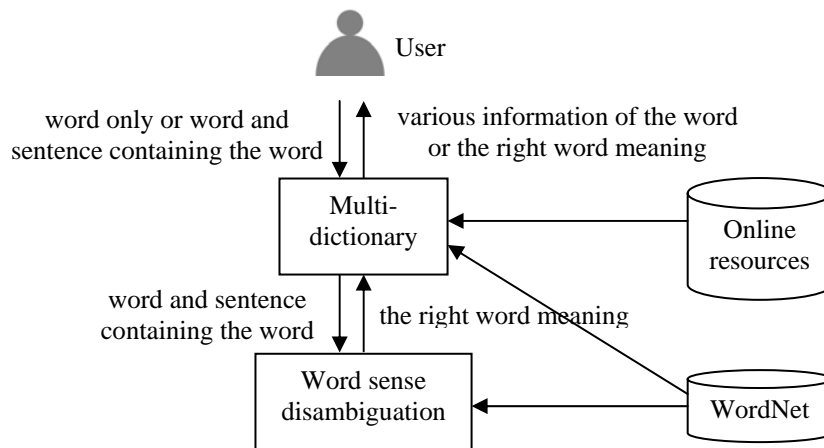


Figure 1. The system architecture of meta-dictionary with word sense disambiguation system

Multi-Dictionary module

The multi-dictionary system begins with query generation from the input word from the user. The next process is to retrieve the input word information such as definition, example of usage, part of speech and so on from online resources using the generated query. The result is a series of web pages containing both desired information about a word as mentioned and other information for example advertisements. The next step is extraction of the desired information and then arranging the material in order using templates to present it to the user.

Word Sense Disambiguation Module

Based on Liu et al. (2005), we employ WSD method with the knowledge base called *WordNet* to our system.

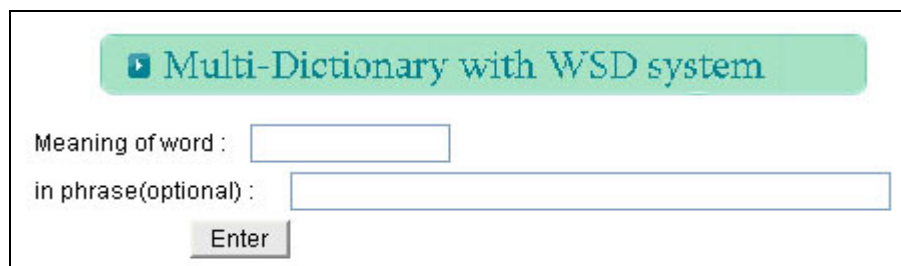
The initial idea of *WordNet* searches dictionaries conceptually, instead of trying to search dictionaries alphabetically. It is an *on-line lexical reference system* (Miller 1990) where English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying *lexical concept* (Miller 1990) and contains words in five categories - nouns, verbs, adjectives, adverbs and function verbs. *WordNet* provides seven kinds of information -- synonyms, definitions, domains, examples, hyponyms, hyponym definitions, and hyponym examples. The lexical information is organized in terms of word meanings, rather than word forms. Therefore, semantic relations are the basis for organization. These relations that include:

- *Synonymy* - a semantic relation between a word meaning the same or nearly the same meaning as another word or other words in a language (answer.com 2006).
- *Antonymy* - a semantic relation between word meanings opposite to that of another word.
- *Hyponymy/Hypernymy* - a semantic relation between word meanings. It is also called as subordination/ superordination, subset/superset, or the ISA relation. X is said to be a *hyponymy* of y if native speakers of English accept the sentence constructed as "An x is a (kind of) y." (Miller 1990) for example "cat" is a hyponym of "animal".
- *Meronymy/Holonymy* - a semantic relation that holds between a part and the whole (Miller 1990). X is said to be meronym of y, if x is a part of y and y is a holonym of x, and vice versa. For example "wheel" is a meronym of "car".

The *WSD* module begins with determining noun phrases of the input context. Next, there are three steps in the disambiguation process. The first step is to employ *WordNet* to gather information such as synonyms, definitions, examples, hyponym but only four pieces of information are significant. The meaning can be determined by the comparison of such information by counting the number of common words between the input word and the noun phrases in the given context. However, if the first step cannot determine the right sense of a word, then a guessing process is conducted in next

step. The second step is to calculate the frequencies of use of the synsets; set of synonym, supplied by *WordNet* to make a guess of the senses of the input words. The higher the frequency of a particular use of word w is, the more likely that this sense is used. Suppose the frequency sum of all senses of w is x . Based on Liu et al. (2005), the threshold is set to a half of x . Using the first sense without any additional information has at least 50% chance of being right. The first sense is called a “dominant sense” of the given term. If no sense of the word has a 50% or higher chance of being used, then the third step is conducted. Finally, for undetermined word apply a web search such as *Google* for the correct sense. The query is submitted to *Google* and the top 20 documents are retrieved. For each document, this method finds a window of y words. This window contains all query terms. Then all the content words in the window are used to form a vector, so that there are 20 vectors, which represent 20 retrieved documents and are combined to form to one vector V . The definition of each sense of the term also forms a vector. The sense of the term whose vector has the highest similarity with V is the determined sense of w . For more information about the systems used, refer to Liu et al. (2005).

The user interface of Multi-Dictionary with *WSD* is shown in Figure 2. The output of the system is selected information from online dictionaries and encyclopaedias or the correct meaning corresponding to the input context. An example of “Robustness” output from word querying is shown in Figure 3 and from word and context “Robustness is a necessary attribute of computer vision” querying is shown in Figure 4.



Multi-Dictionary with WSD system

Meaning of word :

in phrase(optional) :

Enter

Figure 2. The user interface of Multi-Dictionary with Word Sense Disambiguation (WSD) system

▶ Multi-Dictionary with WSD system

Meaning of word :

in phrase(optional) :

▶ Wikipedia

Robustness *From Wikipedia, the free encyclopedia*

In the context of computer software, robustness is the resilience of the system, especially when under stress or when confronted with invalid input. For example, an operating system is considered robust if it operates correctly when it is starved of memory or storage space, or when confronted with an application that has bugs or is behaving in an illegal fashion - such as trying to access memory or storage belonging to other tasks in a multitasking system.

Most modern computer designs have memory protection hardware allowing processes to be forcibly confined to their own memory space. In older designs, such as most 8-bit systems and many early 16-bit ones, this was not available, and thus system integrity was preserved mainly by clean design and careful coding. Thus the perceived robustness of a system became a major factor in debates about different machines and operating systems' quality and performance.

It is the ability of the software system to maintain function even with the changes in internal structure or external environment

Robustness is also a term used to define the ability of a financial trading system to remain effective under different markets and different market conditions.

▶ WordNet

Noun

* S: (n) robustness, hardness, lustiness (the property of being strong and healthy in constitution)

* S: (n) robustness (the characteristic of being strong enough to withstand intellectual challenge) "the lack of robustness in the findings may be due to the small size of the sample"

▶ Dictionary.com

ro-bust
adj.

1. Full of health and strength; vigorous.
2. Powerfully built; sturdy. See Synonyms at healthy.
3. Requiring or suited to physical strength or endurance: robust labor.
4. Rough or crude; boisterous: a robust tale.
5. Marked by richness and fullness; full-bodied: a robust wine.

robust'ly adv.
robust'ness n.

Source: The American Heritage © Dictionary of the English Language, Fourth Edition
Copyright © 2000 by Houghton Mifflin Company.
Published by Houghton Mifflin Company. All rights reserved.

Figure 3. An example of output from the query word “robustness” using Wikipedia, WordNet , other online resources included e.g., Longman, Oxford, and so on

▶ Multi-Dictionary with WSD system

Meaning of word :

in phrase(optional) :

▶ WordNet

Robustness *meaning in Robustness is a necessary attribute of computer vision.*

* S: (n) robustness (the characteristic of being strong enough to withstand intellectual challenge) "the lack of robustness in the findings may be due to the small size of the sample"

▶

Figure 4. An example of the query word “robustness” in a sentence “Robustness is a necessary attribute of computer vision”

Experiment

In this paper, we perform a preliminary experiment to evaluate the performance of *WordNet* in selecting the correct meaning. Our system is developed by using *java*, *HTML* and *perl*. The process begins with acquiring word information programmed using *java* language. After acquiring the information, the system generates web pages in *HTML* using pattern-based programmed in *perl* language. In the meaning selection process, we developed by using *perl* language.

We use two sample sets; Computer Science and Physics sample sets as shown in Table 1 in this evaluation. Each sample set contains 20 phrases. The sample phrases are collected from news articles, textbooks, online resources, journals and magazines. Each phrase contains a multiple meaning term, which we focus on, identified by using bold style.

Table 1. Sample sets used in the preliminary experiment

Computer science	Physics
1. Increasingly complex programs, using symbolic representations of the machine translation	1. In altering current machine, the armature is sometimes stationary.
2. Encapsulation, polymorphism and inheritance	2. There are three external factors that influence the resistance in a conductor .
3. Modifying a variable with volatile tells the compiler	3. There is no displacement for her motion.
4. A transient field is one that does not affect the state of an object.	4. This is not evidence that it has mass since momentum can exist without mass.
5. It has more components that an on/off switch and must be carefully assembled .	5. Have you ever experienced inertia in an automobile while it is braking to a stop?
6. Cases have transparent panels and low heat internal lights to show off the electronics.	6. Since 1942, Dielectric has supplied broadcast equipment, pressurization products
7. The alternative is to buy a very, very large screen with lower resolution	7. Beginning with the quantization of angular momentum
8. The alternative is to buy a very, very large screen with lower resolution	8. It is unique in providing the means to master gauge field theory prior to the advanced study of quantum mechanics.
9. A computer program can be decomposed into a set of independent "threads".	9. The spectra of linear perturbations, including perturbations above nontrivial ground states.
10. The physical connectors and programming interface had to build on	10. The role of intentions as saddle points of Euclidean functional integral and related topics.
11. A virus has to have the chance to execute its code.	11. Those components deteriorated over time.
12. The former is like the propulsion unit of a missile	12. Regrowing bone requires a scaffold that is stiff, long-lasting and safe.
13. Opening an infected file.	13. just doesn't have the stiffness you want,
14. The ability to classify previously unknown words into existing clusters .	14. the scientists analyzed sediments stuck to the shells
15. A first step a complete implementation of basic finite-state operations	15. as the remnants of stromatolites.
16. Two potential sources on the intermediate level	16. These distinctive, layered mounds are the result of colonies of cyanobacteria
17. it was not picked up by mainstream linguists	17. But some researchers have challenged that theory
18. They would be much better than the shallow approaches.	18. chemical activity around hydrothermal vents could produce similar structures
19. kernel based methods such as Support vector machines	19. The stromatolites rank among the oldest signs of life on the earth.
20. Support vector machines have shown superior performance in Supervised learning.	20. to survive in rocks accessible at the earth's surface today.

The accuracy of meaning selection is determined by experts. After each meaning selection, experts consider carefully whether the output is corresponding to the context. The per cent of accuracy is calculated from the frequency of the accurate outputs. The accuracy of meaning selection module for the two sample sets is shown in Table 2. The accuracy of meaning selection of Sample set 1 and Sample set 2 are 85.0% and 80.0%, respectively. The average of accuracy is 82.5%.

Table 2. The accuracy of meaning selection from two sample sets

	Sample set 1: Computer science	Sample set 2: Physics	Average
Accuracy	85.0%	80.0%	82.5%



Concluding remarks

The multi-dictionary with WSD system has the capabilities of collecting various information about a word from online resources such as dictionary.com (2006), Oxford Advanced Learner's Dictionary (OALD) (2006), Longman Dictionary of Comprehensive English (LDOCE) (2006), Cambridge Advanced Learner's Dictionary (CALD) (2006), Wikipedia (2006), *WordNet* (2006). The system then selects the closest meaning according to the context of the word using *WordNet* as a knowledge base. The system differs from conventional on-line dictionaries in having the capability of selecting the most appropriate meaning of a word being used.

The purpose of the system is to facilitate students in studying science in English. The benefit of searching several dictionaries is the variety of descriptions or explanations available for perusal. Students also improve their understanding of discipline vocabulary. The results of the experiment described in this paper demonstrate that the system selects the correct meaning of the word on average in only 82.5% of instances. To improve this outcome we plan to adapt probabilistic methods of natural language processing for the system to select meanings. We also plan to determine student attitudes towards using the multi-dictionary with WSD, and will conduct experiments to assess any increase in student learning as a result of using the system.

Acknowledgement

The authors gratefully acknowledge funding provided by the Institute for Promotion of Teaching Science and Technology (IPST), Thailand.

References

- Answers.com*. [Online] (2006) Available: <http://www.answers.com/synonym> [2006, June 14].
- Cambridge Advanced Learner's Dictionary (CALD)* [Online] (2006) Available: <http://dictionary.cambridge.org/define.asp?key=23046&dict=CALD> [2006, June 14].
- Ide, N. and Veronis, J. (1998) Introduction to the special issue on word sense disambiguation: The state of the art. *Computational linguistics*, **24**(1), 1–40.
- Liu, S., Yu, C. and Meng, W. (2005) Recognition and classification of noun phrases in queries for effective retrieval. *Proceeding of the Conference on Information and Knowledge Management*. Bremen, Germany.
- Longman online dictionary* [Online] (2006) Available: <http://www.ldoceonline.com/> [2005, June 14].
- Miller, G.A. (1990) Wordnet: An on-line lexical database. *International Journal of Lexicography*, **3**(4), 235–312.
- Onelook.com*. [Online] (2006) Available: <http://www.onelook.com> [2005, June 14].
- Oxford Advanced Learner's Dictionary (OALD)* [Online] (2006) Available: http://www.oup.com/oald-bin/web_getald7/index1a.pl [2006, June 14].
- Wikipedia, the free encyclopedia*. [Online] (2006) Available: http://en.wikipedia.org/wiki/Main_Page [2005, June 14].
- WordNet Search 2.1* [Online] (2006) Available: <http://wordnet.princeton.edu/perl/webwn> [2006, June 14].

© 2006 Numtip Rattanawongchaiya, Kanlaya Naruedomkul and Nick Cercone

The authors assign to Uniserve Science and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to UniServe Science to publish this document on the Web (prime site and mirrors) and in printed form within the UniServe Science 2006 Conference proceedings. Any other usage is prohibited without the express permission of the authors. UniServe Science reserved the right to undertake editorial changes in regard to formatting, length of paper and consistency.