Experiments in Topic Detection

Yllias Chali

Department of Mathematics and Computer Science
University of Lethbridge
4401 University Drive
Lethbridge, AB T1K 3M4, Canada
E-mail: chali@cs.uleth.ca

Abstract

Dividing documents into topically-coherent units and discovering their topic might have many uses. We present a system that proceeds in two steps: (1) the input text is segmented at places where there is a probable topic shift, (2) lexical chains are extracted from each segment as indicators of its topic. Two implementations, based on public domain resources, are presented: one based on WordNet and the second one based on Roget's thesaurus. An evaluation of the algorithm shows that lexical chains are acceptable as topic indicator with 44.5% of precision and 63.8% of recall.

1. Introduction

Dividing documents into topically-coherent units and discovering their topic might have many uses. (1) In information retrieval, documents in many collections likely address multiple topics and various aspects of the primary topic. Indexing and clustering these documents based on topical words, instead of frequent phrases, can be exploited to improve the accuracy of an information retrieval system. (2) In text summarization, the primary problem is detecting the relevant portions of texts. Characterizing those portions by their topic will improve the summarization task, especially when the purpose of the summary is user-focused (Mani and Maybury, 1999). (3) In text understanding, the scope of several phenomena is intersentential, the topic can take account of such a scope and hence can help in their resolution, e.g., in resolving anaphora and ellipsis (Kozima, 1993). (4) In structuring text with regard to its discourse hierarchy (Halliday and Hasan, 1976; Hahn, 1990; Morris and Hirst, 1991). (5) In improving document navigation and hypertext links (Green, 1997; Pratt et al., 1999).

Much research has been devoted to the task of structuring text - that is dividing texts into units based on information within the text. Existing work falls roughly into one of the two categories: linear text segmentation aims to discover the topic boundaries, and discourse segmentation focuses on identifying relations between utterances. Methods for finding the topic boundaries include word repetition within a sliding window (Hearst, 1997), lexical cohesion based on word similarity (Morris and Hirst, 1991; Kozima, 1993), entity repetition with regard to its position within the paragraph (Kan et al., 1998), word frequency algorithm and maximum entropy model (Reynar, 1999), context vectors (Kaufmann, 1999), feature induction model (Beeferman et al., 1999), divisive clustering (Choi, 2000). On the other hand, discourse segmentation is fined-grained, (Litman and Passonneau, 1995) combine multiple knowledge sources for discourse segmentation using decision trees, and (Marcu, 2000) uses rhetorical parsing and decision tree to build up the discourse structure based on relations.

The approach that we are proposing to pursue below is a step further to the approaches intending to identify the boundaries between paragraphs in a text where the text changes topic. We present a system that proceeds in two steps: (1) the input text is segmented at places where there is a probable topic shift using one of the following public domain segmenters: *TextTiling* system (Hearst, 1997), *Segmenter* system (Kan et al., 1998), or Choi's system (Choi, 2000), (2) lexical chains are extracted from each segment, using either WordNet or Roget's thesaurus, as indicators of its topic. In this paper, we will present each of these steps, then we will evaluate the whole system. We describe an algorithm for identifying the topic of unrestricted texts. The algorithm takes as input segments of text that represent grouping of contiguous portions of the text, and discovers lexical chains as indicator of their topics. Two implementations, based on public domain resources, are presented: one based on WordNet and the second one based on Roget's thesaurus. The evaluation of the algorithm shows that lexical chains are acceptable as topic indicator with 44.5% of precision and 63.8% of recall, using WordNet.

2. System Overview

The overall architecture of the system is shown in *Figure 1.* It consists of two independent modules organized as a pipeline.

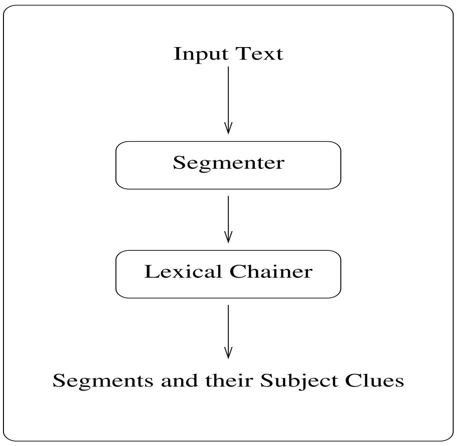

Figure 1: System Overview

3. Text Segmenting

The linear segmentation task is motivated by the observation that comprehension of longer texts benefits from automatic chunking of cohesive sections. This task involves breaking input text into segments that represent some meaningful grouping of contiguous portions of the text. The input text is divided into a linear sequence of adjacent segments and segment boundaries are found at various paragraph separations which identify one or more subtopical shifts.

Multi-paragraph subtopic segmentation should be useful for many text analysis tasks, including information retrieval and summarization, especially, text segmentation is interesting for the following purposes:

- Segmentation is intended to identify the boundaries between paragraphs in a text where the text changes topic. Thus, a text can comprise merely a single segment, or perhaps several different segments, when it touches on several different topics.

- It helps in processing the user needs when they are specified as terms in the sense that only segments that are relevant to the terms specified by the user are chosen (Reynar, 1999; Chali et al., 1999). When the topically-coherent units (i.e., text segments) are represented by a set of topical clues, a content-based process of matching the user's terms against the segment's clues will determine the relevancy of the segments, i.e., the segments with the highest matches are selected as answers to the user's query.

We are using three public domain segmenters *TextTiling* system (Hearst, 1997), *Segmenter* system (Kan et al., 1998), and Choi's system (Choi, 2000).

Segmentation is followed by the characterization of the segment in terms of lexical chains as clues of the segment topic.

4. Lexical Chaining

Structural theories of text are concerned with identifying units of text that are about the "same thing". When this happens, there is a strong tendency for semantically related words to be used within that unit. The notion of cohesion, introduced by (Halliday and Hasan, 1976), is a device for "sticking together" different parts of the text to function as a whole. It is achieved through the use of *grammatical cohesion*, i.e., reference, substitution, ellipsis and conjunction, and *lexical cohesion*, i.e., semantically related words. Lexical cohesion occurs not only between two terms, but among sequences of related words, called *lexical chains* (Morris and Hirst, 1991). Lexical chains (1) provide an easy-to-determine context to aid in the resolution of ambiguity and in the narrowing to specific meaning of a word, (2) tend to delineate portions of text that have a strong unity of meaning. We investigate how lexical chains can be used as an indicator of the text segment topic. The steps of the algorithm of the lexical chain computation are as follow:

1. We select the set of candidate words. To this end, we run a part-of-speech tagger (Brill, 1992) on a text segment, and only the open class words that function as noun phrases or proper names are chosen.

2. The set of the candidate words are exploded into senses, the senses are given by the thesaurus in use, at this step all the senses of the same word are considered. In the actual implementation, we are using two different thesauri: Roget's thesaurus (Chapman, 1988) and WordNet thesaurus (Miller et al., 1993). From this step each word sense is represented by distinct sets (see *Figure 2*) considered as levels, the first one constitutes the set of synonyms and antonyms, the second one constitutes the set of first hypernyms/hyponyms and their variations (i.e. meronyms/holonyms, etc.), and so on.

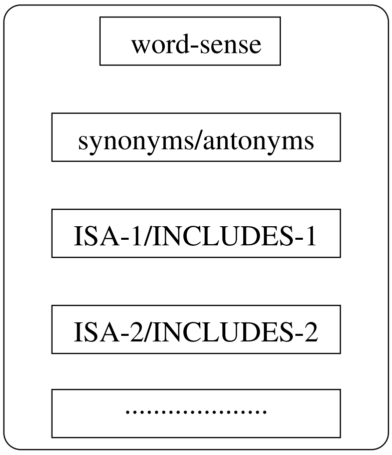

Figure 2: Word sense representation

3. We find the semantic relatedness among the set of senses according to their representations. A semantic relationship exists between two word senses if comparing two sense representation of two distinct words, a matching exists, i.e., a non-empty intersection exists between the sets of words. To each semantic relatedness is associated a measure which indicates the length of the path taken in the matching with respect of the levels of the compared two sets.

4. We build up chains which are sets such as

$$\begin{array}{ll} \{ & (word_1[sense_{11}, sense_{12}, \ldots]), \\ \{ & (word_2[sense_{21}, sense_{22}, \ldots]), \\ \{ & \ldots\} \end{array}$$

in which $word_\imath$-$sense_{\imath x}$ is semantically related to $word_\jmath$-$sense_{\jmath y}$ for $\imath \neq \jmath$, and x and y correspond to the senses of $word_\imath$ and $word_\jmath$, respectively.

5. We retain longest chains relying on the following preference criterion:

$$\begin{array}{l} word\ repetition\ \gg \\ \quad synonym/antonym\ \gg \\ \qquad ISA-1/INCLUDE-1\ \gg \\ \qquad\quad ISA-2/INCLUDE-2\ \gg \\ \qquad\qquad \ldots \end{array}$$

> In our implementation, this preference is handled by assigning scores to each pairwise of semantical relatedness in the chain, and then adding up those pairwise scores. Hence, the score of a chain is based on its length and on the type of relationships holding among its members.

In the lexical chaining method, the relationship between the words in a chain is pairwise mutual, that is, each word-sense has to be semantically related to every other word-senses in the chain. The order of the open class words in the document does not play a role in building up the chains. However, it turned out that the number of lexical chains could be extremely large, and thus problematic, for larger segments of text. To cope with, we reduced the word-sense representation to synonyms only, when we have long text segments. This reduction has another benefit, in the sense that a lexical chain based only on synonyms could be better than one based on ISA-2/INCLUDE-2. This reduction also has to narrow down the set of lexical chains stemming from the single segment in the case when they are too many.

We will show the output of the lexical chaining on a fragment of text (1), (2) and (3) are the lexical chains computed using WordNet and Roget's thesauri, respectively.

(1) A series of explosions and fire shut down electricity generation at the world's largest solar power plant near here Wednesday. Thick plumes of black smoke spiraled into the clear desert air when one of four natural gas-fired heaters used to back up the solar heating system exploded. A short time later, a second natural gas heater caught fire and exploded as the first of 75 firefighters and 25 pieces of equipment were arriving at the site, about 140 miles northeast of Los Angeles. "We had a series of explosions, more than two," said Capt. Sharon Sellers of the San Bernardino County Fire Department. "Our first units got on-scene at 9:16 a.m. and a second explosion occurred at that point, then a series of them during the entire incident," Sellers said. "There was a mushroom cloud. The heat was real intense and there were explosions," said an inmate from the Boron Federal Prison Camp who was pressed into service to help fight the fire. He would not identify himself. Sellers said two workers at the plant suffered minor breathing problems and were treated at Barstow Community Hospital. Operated by LUZ International Ltd. of Los Angeles, the $280-million Harper Lake solar plant began generating electricity on Dec. 28 and produces 80 megawatts, enough power to serve 115,000 people. The company operates eight such plants in the California desert. Combined, they generate 274 megawatts, which is sold to Southern California Edison Co. An Edison spokesman said there was no interruption of electric service to its customers. "We had two oil heaters on line and were bringing up the third and fourth oil heaters when this explosion occurred," LUZ International spokeswoman Kathleen Flanagan said in Los Angeles. While no flames were visible 1 1/2 hours after the fire began shortly before 9 a.m., San Bernardino County firefighters had difficulty reaching the blaze deep within the generating equipment. "There is fire up there somewhere still heating that oil," Sellers said. The blaze was contained, but continued to burn late Wednesday. Cause of the fire was unknown, but fire officials ruled out arson and said it probably resulted from an equipment malfunction. While Flanagan said she could not immediately estimate the cost of the blaze, the Fire Department said a single natural gas heater costs $500,000. One was destroyed and a second was heavily damaged. Flanagan said the black smoke from an estimated 15,000 gallons of burning synthetic oil was not any more toxic than smoke from natural crude or refined oil and was not carcinogenic. But that report was disputed by Capt. Clyde Gamma of the California Department of Forestry and Fire Protection. He identified the synthetic oil as Therminol and said it is cancer-causing. Flanagan said the plant could resume generating electricity by Monday. But she said the backup natural gas-fired heaters would not be used. "We will be operating strictly in the solar mode," she said. For solar generation, large curved mirrors are used to concentrate the sun's energy onto synthetic oil, which flows through an insulated steel pipe. The hot oil boils water into steam that drives conventional electrical turbines. Sellers said LUZ International had a fire about two years ago at another solar plant at Daggett and that explosions continued five hours into the incident. Stammer reported from Los Angeles and Harris from Barstow.

(2) a. { blaze , fire }
b. { breathing, heater, smoke }
c. { california, los_angeles }
d. { company, ltd. }
e. { county, department }
f. { crude_oil, oil }
g. { desert }
h. { difficulty, problem }
i. { electricity }
j. { equipment, mode }
k. { equipment, unit }
l. { explosion, fire }
m. { fire, protection }
n. { gas_heater, heater, oil_heater, smoke }
o. { international }
p. { monday, wednesday }
q. { plant, worker }
r. { firefighters }
s. { barstow }
t. { luz }
u. { flanagan }
v. { generating }

(3) a. { air, difficulty, line }
b. { air, line, pipe, series, unit }
c. { air, line, mode }
d. { air, line, piece, report }
e. { cause, energy }
f. { cause, world_power }
g. { county, department }

h. { fire, protection }
i. { international }
j. { monday, wednesday }
k. { cloud, electricity, energy, power }
l. { cloud, mushroom }
m. { company, system, units }
n. { difficulty, problems }
o. { energy, heating }
p. { plant, worker}
q. { firefighters }
r. { barstow }
s. { luz }
t. { flanagan }

Lexical chains are computed for each text segment. They are sets of clues reflecting the topic of the text segment.

5. Evaluation

We conducted an evaluation of the whole system (i.e., segmenter + lexical chainer). We selected randomly ten texts from the Brown corpus as test corpus, we segmented them using Choi's segmenter because it is more precise than the two others (cf. (Choi, 2000)), this gave us a sample of 112 text segments, then we computed the lexical chains for each of these segments using both of the thesauri, and we presented them to five judges. We asked all the judges:

i. to read the text segment, then

ii. to read each lexical chain and answer the following question:

$$“Is\ the\ chain's\ topic\ present\ in\ the\ segment?'' \quad (1)$$

iii. after reading all the lexical chains corresponding to one segment and answering the previous question, we asked them to answer the following question:

$$“Is\ the\ segment's\ topic\ covered\ by\ all\ its\ chains?'' \quad (2)$$

We considered the answer as *yes* or *no* given the majority of the judges. Related to the information retrieval measures, the answers to the first question correspond to precision (i.e., how many lexical chains are good among all the computed lexical chains), the answers to the second question correspond to recall (i.e., how much of the segment's topic is contained in the lexical chains). Precision and recall are computed according to Equations (3) and (4), respectively.

$$Precision = \frac{number\ of\ answer\ yes\ to\ question\ 1}{total\ number\ of\ question\ 1} \quad (3)$$

$$Recall = \frac{number\ of\ answer\ yes\ to\ question\ 2}{total\ number\ of\ question\ 2} \quad (4)$$

We notice that $total\ number\ of\ question\ 1$ corresponds to the total number of chains, and $total\ number\ of\ question\ 2$ corresponds to the total number of segments.

The results of our evaluation are shown in *Table 1.*

	Precision	Recall
Using WordNet	44.5%	63.8%
Using Roget's thesaurus	38.7%	54.6%

Table 1: Results of the evaluation

This experiment shows that the whole system is more accurate using WordNet than Roget's thesaurus. This is due on one hand to the number of entries in the thesaurus (i.e., 99,642 synsets and 121,962 unique words in WordNet as of version 1.6 compare to Roget's thesaurus 1035 categories and 46,500 unique words as of version 7.1). On the other hand, the classification into categories in Roget's thesaurus are more general abstraction compare to the organization into synsets defined in WordNet. Indeed, WordNet represents the largest publically available lexical resource to date.

6. Related Work

The goal of text categorization is to learn a classification scheme that can be used for the problem of automatically assigning arbitrary documents to predefined categories or classes. There has been a wide range of statistical learning algorithms applied to this automatic text categorization task. They include the Rocchio relevance feedback algorithm (Buckley et al., 1994; Joachims, 1997; Lewis et al., 1996), k-Nearest Neighbor classification (Yang, 1994), naive Bayes probabilistic classification (Joachims, 1997; Lewis and Ringuette, 1994; McCallum and Nigam, 1998), support vector machines (Joachims, 1998), and neural networks (Wiener et al., 1995). After two preprocessing steps: representation and feature selection, answering the questions of how to discriminate informative words in the reduced vector space and how to give them more weight than other non-informative words is the main task of classifiers. This approach relies on a mapping from a new document to relevant categories, given we have categories that are built manually. Our approach does not need training and as a consequence can be used when no a priori hypothesis can be done about topics that are concerned.

Other systems are based on world knowledge. For instance, (DeJong, 1982) developed a system based on templates that organize its world knowledge in order to skim newspaper stories and extract the main details. (Radev and McKeown, 1998) developed a system that takes template outputs of information extraction systems developed for the MUC conference and generates summaries of multiple news articles. Those systems rely on prior knowledge of their domains. However, to acquire such prior knowledge is labor-intensive and time-consuming. In order to reduce the knowledge engineering bottleneck, (Riloff and Lorenzen, 1999) present a system that generates extraction patterns and learns lexical constraints automatically from preclassified texts. (Lin and Hovy, 2000) present a procedure to automatically acquire topic signatures from preclassified documents of specific topics which are then used to identify the presence of the learned topics in previously unseen documents. However, learning extraction patterns from corpora makes those systems domain-specific. We presented a method based on common available resources such as WordNet, and which can be applicable for unrestricted texts.

Lexical chains has been proposed by (Morris and Hirst, 1991) as indicator of the structure of text. (Barzilay and Elhadad, 1997) investigate the production of summaries based on lexical chaining. The summaries are built using scoring which is based on chain length and the extraction of significant sentences is based on heuristics using chain distribution, for example, choose the sentence that contains the first appearance of a chain member in the text. (Yarowsky, 1992) presented statistical models using lexical chains for the purpose of word-sense disambiguation. (Ellman, 2000) uses the lexical chains to determine the similarity of texts. In this paper, we investigated the production of lexical chains to account for the topic of the text segment.

The described algorithm for the lexical chaining was implemented in C++. Its primary purpose is to extract from the text segments meaningful clues as indicator of the segment's topic. This technique has many uses in processing and searching of information.

The results reported in this paper suggest that we may refine the process of lexical chaining. Instead of choosing any content word tagged as noun or proper noun as candidate for the computation of the chains, it seems that restricting the set of candidate words will improve the precision of the chains.

7. Conclusion

Topic detection and identification is an important area of research, addressing many application needs. It presents new and interesting technical challenges.

We presented an algorithm for detecting the topic of unrestricted texts based on an efficient use of lexical chains acquired from common lexical knowledge. The results show that the algorithm is promising for many applications where an efficient access to large quantities of information is needed.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) research grant and the Alberta Heritage Foundation for Science and Engineering Research under the Research Excellence Envelope funding.

8. References

R. Barzilay and M. Elhadad. (1997). Using lexical chains for text summarization. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 10-17, Madrid.

D. Beeferman, A. Berger, and J. Lafferty. (1999). Statistical models for text segmentation. *Machine Learning, Special Issue on Natural Language Processing*, 34(1 - 3):177 - 210.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*, pages 152-155, Trento.

C. Buckley, G. Salton, and J. Allan. (1994). The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292-300.

Y. Chali, S. Matwin, and S. Szpakowicz. (1999). Query-biased text summarization as a question-answering technique. In *Proceedings of AAAI Symposium on Question-Answering Systems*, pages 52-56, Massachusetts. AAAI Press.

R. Chapman. (1988). *Roget's International Thesaurus*. Longman, London.

F. Y. Y. Choi. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics*, pages 26 - 33, Seattle, Washington.

G. DeJong. (1982). An overview of the frump system. In Wendy G. Lehnert amd Martin H. Ringle, editor, *Strategies for natural language processing*, pages pages 76-49. Lawrence Erlbaum Associates.

J. Ellman. (2000). *Using Roget's Thesaurus to Determine the Similarity of Texts*. Ph.D. thesis, School of Computing Engineering and Technology, University of Sunderland.

S. J. Green. (1997). *Automatically Generating Hypertext by Computing Semantic Similarity*. Ph.D. thesis, Department of Computer Science, University of Toronto.

U. Hahn. (1990). Topic parsing: Accounting for text macrostructures in full text analysis. *Information Processing and Management*, 26:135 - 170.

M. Halliday and R. Hasan. (1976). *Cohesion in English*. Longman, London.

M. A. Hearst. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33-64.

T. Joachims. (1997). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 143-151.

T. Joachims. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*.

M. Y. Kan, K. R. McKeown, and J. L. Klavans. (1998). Linear segmentation and segment relevance. In *Proceedings of 6th International Workshop of Very Large Corpora (WVLC-6)*, pages 197-205, Montréal.

S. Kaufmann. (1999). Cohesion and collocation: Using context vectors in text segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (Students Session)*, pages 591 - 595, College Park, Maryland.

H. Kozima. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 286-288.

D. D. Lewis and M. Ringuette. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the 3rd Symposium on Document Analysis and Information Retrieval*.

D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka. (1996). Training algorithm for linear text classifiers. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298-306.

C. Y. Lin and E. Hovy. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of 18th International Conference in Computational Linguistics*, Saarbrücken, Germany.

D. J. Litman and R. J. Passonneau. (1995). Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 108 - 115, Cambridge, Massachusetts.

I. Mani and M. Maybury. (1999). *Advances in Automatic Text Summarization*. MIT Press.

D. Marcu. (2000). *Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

A. McCallum and K. Nigam. (1998). A comparison of event models for naive bayes text classifiers. In *Proceedings of AAAI Workshop on Learning for Text Categorization*.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. (1993). Five papers on wordnet. CSL Report 43, Cognitive Science Laboratory, Princeton University.

J. Morris and G. Hirst. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21-48.

W. Pratt, M. A. Hearst, and L. M. Fagan. (1999). A knowledge-based approach to organizing retrieved documents. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, Orlando, Florida.

D. R. Radev and K. R. McKeown. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469-500.

J. C. Reynar. (1999). Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 357 - 364, College Park, Maryland.

E. Riloff and J. Lorenzen, (1999). *Generating domain-specific role relationships automatically*, chapter Natural Language Information Retrieval, Tomek Strzalkowski editor. Kluwer Academic Publishers.

E. Wiener, J. O. Pederson, and A. S. Weigend. (1995). A neural network approach to topic spotting. In *Proceedings of the 4th Symposium on Document Analysis and Information Retrieval*.

Y. Yang. (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 13-22.

D. Yarowsky. (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.