Intl. Conf. RIVF'04
February 2-5, Hanoi, Vietnam

Lexical Similarity based on Quantity of Information Exchanged - Synonym Extraction

Ngoc-Diep Ho, Fairon Cédrick

***Abstract*—There are a lot of approaches for measuring semantic similarities between words. This paper proposes a new method based on the analysis of a monolingual dictionary. We can view the word definitions of a dictionary as a network: its *nodes* are the headwords found in the dictionary and its *edges* represent the relations between a headword and the words present in its definition. In this view, the meaning of a word is defined by the *total quantity of information*, in which each element of its definition contributes. The similarity between two words is defined by the maximal *quantity of information exchanged* between them through the network.**

In order to assess the performance, our measure of similarity will be compared with others measures and some applications using this measure will also be described.

***Keywords*—Lexical Similarity, Synonym Extraction, Information Exchanged**

I. INTRODUCTION

The long history of formalizing and quantifying semantic similarities between lexical units began at the latest with Aristote (384 – 322 B.C) [1]. We all know that *house*, *apartment* and *flat* share some features and that they can be sometimes used interchangeably. However, only a measure of similarity could tell us how much one word is semantically close to another (i.e. *sim(house, apartment)*, *sim(house,flat)* and *sim(flat,apartment)*). There are a lot of Natural Language Processing (NLP) methods for measuring semantic similarity between words, which are based on different approaches. In the following sections, we will propose a new method based on the analyses of a monolingual dictionary.

We can view the definitions of words in a dictionary as a network. Its vertices are the headwords found in the dictionary and its edges represent relations between a headword and the words found in its definitions. In this view, the meaning of a word is defined by the total quantity of information to which each element of its definition contributes. The similarity between two words is defined by the maximal quantity of information exchanged (QIE) between them through the network.

The approach is easy to adapt to various languages because it requires only a monolingual dictionary. Although the resource we have used is not very structured, the quality of our experimental results is equivalent to other existing measures, including the ones based on more structured resources like WordNet. These results are interesting enough to open new research directions in NLP but they could probably be improved by taking into account additional linguistic aspects (for example, at a morphological or lexical level).

This paper is organized as follows: section II summarizes some existing methods found in the literature, section III describes our method based on the new notion of "Quantity of Information Exchange", section IV presents the application of the method to English, section V presents a synonym extractor based on the results of section IV and finally, section VI presents our conclusions and perspectives.

Ngoc-Diep Ho – Faculty of Applied Mathematics, University of Louvain, Belgium; Email: ho@inma.ucl.ac.be.

Fairon Cédrick – Center for Natural Language Processing (CENTAL), University of Louvain, Belgium; Email: fairon@tedm.ucl.ac.be.

II. RELATED WORKS

In recent years, many researchers have proposed new definitions of lexical similarity. The possibility of quantifying the lexical proximity between words opens ways for the semantic processing of text. The main existing methods can be categorized into several groups:

- Methods that use a monolingual dictionary ([9], [10]…),
- Methods that use WordNet ([5], [11] …),
- Methods that use WordNet and use some analyses on a textual corpus ([7], [15], [12] …),
- Methods that use a thesaurus ([14] …).

Methods that use WordNet or a thesaurus may give very good results, because WordNet and thesauri are manually created and the relations between words in these resources are quite explicit. Therefore, the similarity between two words depends simply on their relative position in the resource. The frequent choice of these methods to evaluate the level of relationship between two words is: calculating the shortest path that connects the words. In addition, a corpus may be used to put some statistics into action such as probability of words, collocations of words, etc.

Performances of the last 3 groups are very good, so why are many researchers, including us, trying to create new methods based solely on a monolingual dictionary (the first approach)? The reason is that WordNet and good thesauri (for example Roger) exist only for a few languages (English, French…), while monolingual dictionaries exist in most languages in the world. Hence, an application that uses the methods of the first

group will be easily adapted to all languages (because fewer language resources will be needed).

III. SIMILARITY BASED ON QUANTITY OF INFORMATION EXCHANGED

The new definition of similarity that we present here is based on the interconnection network of concepts and the informational content of these concepts. In that network, each *vertex* represents a concept and each *edge* represents the relationship between 2 concepts. The informational content of concepts (vertices) and relations (edges) are accounted only in terms of their quantitative aspect (i.e. the quantity of information) but not in term of their qualitative aspect (i.e. the semantic type of information). These initial ideas lead to the following 2 important intuitions.

> **Intuition 1:** the description of a concept is constituted by the quantity of information that its neighbor concepts transfer to it.

In figure 1, suppose that we do not know the concept in the center x whose neighbor concepts are $O_1, O_2, ..., O_n$. But by knowing all its neighbors, we can more or less figure out what the concept x is. Actually, $O_1, O_2, ..., O_n$ have transferred a certain amount of information to x, so that we can have some knowledge of x through $O_1, O_2, ..., O_n$.

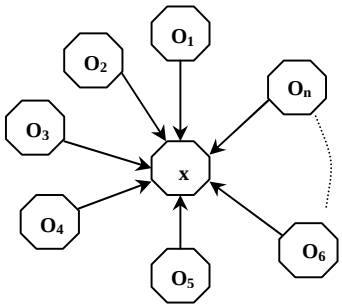

Figure 1: unknown concept and its neighbors

Again, the descriptions of $O_1, O_2, ..., O_n$ are themselves constituted by information transferred from their own neighbors. Hence, by putting all the concepts in an interconnection network, we reach the second intuition.

> **Intuition 2:** in network of concepts, the similarity between the concept A and the concept B is dependent on the quantity of information that A can transfer to B and on the quantity of information that B can transfer to A. In other words, the similarity between A and B is dependent on the quantity of information exchanged between A and B through a *network of concepts*.

Now we need to formalize these ideas. First, the informational content of a concept can be calculated by the well-known formula from the theory of information:

$$I(A) = -\log(P(A))$$

where $P(A)$ is the probability of A. A concept A can transfer a fraction of its informational content to its neighbors. And ideally, the quantity of information that A can receive is equal to its informational content.

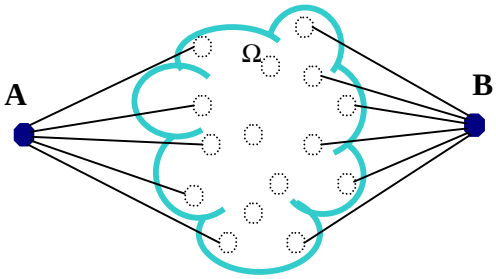

Figure 2: two concepts in a concept network

To calculate the similarity between concepts, we must know the maximal amount of information that each edge in the concept network can hold; we call this amount the capacity of the edge. Unfortunately, this capacity of an edge seems to be dependent on the *similarity* between 2 concepts on that edge. And we can not calculate it because we have no knowledge of this similarity yet. So we try to estimate these capacities, on the sole basis of the informational content of concepts and on the structure of the concept network.

Normally, a concept that has higher informational content can transfer and receive higher amount of information to/from its neighbors. Therefore, we estimate the edges capacities as below:

$$I(O_i, A) = w_i . I(A)$$

(Capacity of the edge from O_i to A),

where $w_i = \dfrac{I(O_i)}{\sum_{O_j \in Neighbors(A)} I(O_j)}$

Thanks to these estimations, we can construct a complete concept network in order to calculate the similarity between concepts. As stated in the intuition 2, the degree of similitude between two concepts A and B in the network Ω is equal to the quantity of information exchanged between A and B through Ω, so we have:

$$sim_{QIE}(A, B) = f(mfi_{\Omega}(A, B), mfi_{\Omega}(B, A))$$

where $mfi_{\Omega}(A, B)$ is the value of maximum flow (of information) from A to B through the network Ω, f is a function that combines the value of maximum flow from A to B and the value of maximum flow from B to A. Two natural choices for this function are:

$$sim_{QIE1}(A,B) = \frac{mfi_{\Omega}(A,B) + mfi_{\Omega}(B,A)}{2}$$

and

$$sim_{QIE2}(A,B) = \sqrt{mfi_{\Omega}(A,B).mfi_{\Omega}(B,A)}$$

We will experiment both possibilities (named below *QIE1* and *QIE2*) in the following sections.

IV. SIMILARITY BETWEEN ENGLISH WORDS

In the previous section, a new definition of lexical similarity has been described. In order to experiment this method on the English language, we will have to create a network of English words on which we can compute the similarity. To do so, we have used the US Webster 1913 dictionary [19], which is freely available online thanks to the Gutenberg project: www.gutenberg.net. The application of our method to more structured and better created resources like WordNet and Roget is also possible, but, as explained above, these kinds of resources are only available for a few languages. Therefore the US Webster is the best candidate in order to show that our method is applicable to many languages.

The available version of the Webster dictionary contains 27 HTML files. Each of the first 26 files contains the definition of all the words that begin with a letter (ranging from A to Z). The 27th file contains the newly added words of the dictionary.

One way to transform this dictionary into a graph is described in [17], [18]. Each headword of Webster is modeled by a vertex in this graph. And an edge is added from the word w_i to the word w_j if w_j is present in the definition of the word w_i. After doing some pre-processing, the resulting graph contains 112 169 vertices and 1 398 121 edges. Several features of this graph were also analyzed in Senellart's report and paper.

All the experiments described in the present paper make use of the same graph as an input. The graph is then converted into a network by making all the edge bi-directional and adding a capacity on each edge, as described in the previous section.

A. QIE Similarity in the Webster dictionary

In our experiments, we used the complete network of all the English words (i.e. all the words in the dictionary) to calculate the similarity. But there were two main drawbacks:

- The algorithm of maximal flow is very time-consuming. The larger the graph network is, the longer it takes to calculate the similarity.
- When the connections between two words are too long, the information which is exchanged between them might not be significant. Semantically, the information exchanged through a long connection will be too general to characterize the level of relationship between words.

Hence, it is wiser to reduce the graph size by selecting only a subgraph that contains all the neighbors of A and B when calculating the similarity between A and B.

Two algorithms of maximum flow were tested: Ford&Fulkerson [2] and Prelow-Push [3]. Although, in theory, the algorithm Preflow-Push has a lower complexity, but in this case, the algorithm of Ford&Fulkerson seems to work faster. A deep analysis of these algorithms is out of the scope of this report.

B. EXPRIMENTAL RESULTS

One way to assess an automatic method aimed at measuring words similarities is to confront the results to the human judgments [1]. Two set of tests were created by Rubenstein&Goodenough [16] (65 pairs of words) and Millers&Charles [13] (30 pairs of words). All the pairs in the 2 sets have been judged by several human subjects and the averages of the given scores of similarity (varying from 0 to 4) were computed.

We have compared our results to both set of tests: first, we have used our method to calculate the similarity of each pair of words given in the lists, and second we have calculated the correlation coefficient between our results and the human judgments. Table 1 presents our results as well as results obtained with other methods.

Methods	Rubenstein-Goodenough	Miller - Charles
Hirst and St-Onge	0.78614403	0.74439909
Jiang and Conrath	0.78127462	0.85002672
Leacock and Chodorow	0.83822965	0.81574130
Lin	0.81930235	0.82917110
Resnik	0.77868458	0.77363821
QIE1	**0.75690479**	**0.75158059**
QIE2	**0.78599606**	**0.83272219**

Table 1: Correlation coefficients of similarity measures

The numerical results show that, despite the use of the poorly structured Webster dictionary, our method has provided very good results. The performance of our method (especially with QIE2), can be considered as equivalent to other methods. Moreover, it is likely that if a better dictionary was used, results would be even better.

V. APPLICATION: SYNONYM EXTRACTOR

Using our new method to measure semantic similarities, we have built a synonym extractor for English (of course, *synonym* must be heard here in a general way). Given an English word *w*, this extractor will try to find *n* words whose meaning is well related to *w*. The results are sorted with respect to the similarity between each word in the list and *w*. The extractor takes the following steps to extract the synonyms of the word *w*:

1. For each word w_k in the dictionary, compute the similarity between w_k and *w*.
2. Sort the list of w_k.
3. Take *m* words that have the greatest similarity with *w* as the synonyms of *w*.

	Distance	Senellart	ArcRank	QIE_L	WordNet
1	Vanish	Vanish	Epidemic	Vanish	Vanish
2	Pass	Pass	Dissapearing	Fade	go away
3	Wear	Die	Port	Wear	End
4	Die	Wear	Dissipate	Die	Finish
5	Light	Faint	Cease	Pass	Terminate
6	Fade	Fade	Eat	Dissipate	Cease
7	Faint	Sail	Gradually	Faint	
8	Port	Light	Instrumental	Light	
9	Absorb	Dissipate	Darkness	Evanesce	
10	Dissipate	Cease	Efface	Disappearing	

Table 2: Synonyms of *Disappear*

	Distance	Senellart	ArcRank	QIE_L	WordNet
1	Cane	Cane	Granulation	Inversion	Sweetening
2	Starch	Starch	Shrub	Dextrose	Sweentener
3	Juice	Sucrose	Sucrose	Sucrose	Carbonhydrate
4	Obtained	Milk	Preserve	Lactose	Saccharide
5	Milk	Sweet	Honeyed	Cane	organic compound
6	Sucrose	Dextrose	Property	Sorghum	Saccarify
7	Molasses	Molasses	Sorghum	Candy	Sweeten
8	Sweet	Juice	Grocer	Grain	Dulcify
9	White	Glucose	Acetate	Root	Edulcorate
10	Plants	Lactose	Saccharine	Starch	Dulcorate

Table 3: Synonyms of *Sugar*

For example, for the word *sugar*, our system provides the following list of synonyms and similarity values: *mucic* (9.66775), *betain* (9.23731), *electuary* (9.21128), *ferments* (8.8084), *muscovado* (8.72424), *chard* (8.68176), *levorotatory* (8.65445), *medicated* (8.62637), *helleborin* (8.60096), *inspissated* (8.56856), *pastille* (8.55456), *massicot* (8.41452), *sizing* (8.40321), *dulcite* (8.39814), *confect* (8.3634), …

By looking up in the Webster 1913 dictionary, we can see that the meaning of these words (except *sizing*) share a lot of features of the word *sugar* and can thus be considered as "synonyms" of *sugar*.

Again, since the algorithm is very time-consuming, we have to narrow the search of synonyms and consider only a list of words that have tight relations (i.e. the length of link is small) with the original word. And the simple choice to obtain this list for the word *w* is to take all vertices in the graph of *neighbors* of *w*.

Tables 2 and 3 contain the "synonyms" of *sugar* and *disappear* automatically provided by different methods (we took these lists in [17]). In these tables, QIE_L stands for the method that computes the QIE similarity only for the neighbors of the word for which to extract synonyms. Because only a limited list of words is taken into account, we may not find all the words that have the greatest similarity. For example, with *sugar*, the following words are not extracted by QIE_L: *mucic*, *betain*, *electuary*, etc. And synonyms with a lower degree of similitude are extracted: *inversion* (7.86811), *dextrose* (7.09966), *sucrose* (6.92324), *lactose* (6.74817), *cane* (6.43), *sorghum* (6.42011), *candy* (6.36874), etc. With a limited number of examples, we can not determine whether our extractor is the better among these methods. But it shows that this extractor can give good synonyms.

A demo of this application is publicly available at: http://cental.fltr.ucl.ac.be/synonyms.

VI. CONCLUSION

A lot of problems are involved in semantic processing of texts. The measure of word sense similarity is only one of them, but it has many possible applications in NLP.

It seems to us that a good method for measuring lexical similarity must be adaptable to various languages and must be, of course, as close as possible to the human judgment. The measure we have proposed, which is based on the quantity of information exchanged (QIE), meet the first criteria and offer promising results regarding the second: on one hand, the use of a simple monolingual dictionary makes our method adaptable to many languages and on the other hand, our experiments on English showed that the method provides reliable results which are largely compatible with the human intuition.

With these interesting results, we still hope to improve them in the future by taking into account more linguistic aspects in the dictionary processing (at a morphological or lexical level, for instance) and create more applications.

REFERENCES

[1] Alexander Budanisky, Lexical Semantic Relatedness and Its Applications in Natural Language Processing, *Rapport Technique* CSRG-390, *Computer Research Group* – University of Toronto.

[2] L. R. Ford and D. R. Fulkerson. Flows in Networks. Princeton Univ. Press, Princeton, NJ, 1962

[3] A. V. Goldberg. A New Max-Flow Algorithm. Technical Report MIT/LCS/TM-291, Laboratory. For Computer Science, MIT, 1985

[4] Graeme Hirst et David St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms. Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press, 1998

[5] Ho Ngoc Diep, Similarité de mots et extraction automatique de synonymes. University of Louvain. Internship Report, Belgium 2002.

[6] Jan Jannink and Gio Wiederhold. Thesaurus Entry Extraction from an On-line Dictionary. In Proceedings of Fusion '99, Sunnyvale CA, July 1999.

[7] Jiang, J.& Conrath, D.W, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of the 10th International Conference: Research on Computational Linguistics (ROCLING X), Academica Sinica, pages 19-33, 1997, Taiwan

[8] J. Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Version extended in Journal of the ACM 46 (1999).

[9] Hideki Kozima, Teiji Furugori. Similarity between words computed by spreading activation on an English Dictionary. Proceedings of EACL-93 (Utrecht), pages 232-239, 1993.

[10] Hideki Kozima and Akira Ito, Context-Sensitive Word Distance by Adaptive Scaling of a Semantic Space. Ruslan Mitkov and Nicolas Nicolov (editors.), Recent Advances in Natural Language Processing (a serie of "Contemporary Issues in Linguistic Theory" 136), pages 111-124, John Benjamins, Amsterdam/Philadelphia, 1997.

[11] Claudia Leacock e[Leacock&Chodorow 98] Claudia Leacock et Martin Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. Christiane Fellbaum (ed.). WordNet: an electronic lexical database. Cambridge: MIT Press, pages 265-283.

[12] D. Lin. An Information-Theoretic Definition of Similarity. In Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July, 1998.

[13] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. In Language and Cognitive Processes, 6(1): pages 1 - 28, 1991.

[14] Manabu Okumura, and Takco Honda, Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. In Proceedings of Fifteenth International Conference on Computational Linguistics (COLINGS-94), vol.2, pages 755-761, Kyoto, Japan, August 1994.

[15] Philip Resnik 1995, Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[16] H. Rubenstein, & J. Goodenough, 1965. Contextual correlates of synonymy. CACM, 8 (10), pages 627--633.

[17] Pierre Senellart 2001. Extraction of information in large graphs – Automatic Search of Synonymes, Rapport de stage, Université Catholique de Louvain.

[18] Pierre P. Senellart, Vincent D. Blondel, Automatic discovery of similar words, chapter in: Survey of Text Mining, Springer-Verlag, 2003.

[19] The Online Plain Text English Dictionary, http://msowww.anu.edu.au/~ralph/OPTED/, in the project of Gutenberg 2000.