Semantic Matching Using Concept Lattice

Ana Meštrović

University of Rijeka, Department of Informatics,
Omladinska 14, 51000 Rijeka, Croatia
amestrovic@inf.uniri.hr

Abstract. This paper describes how a concept lattice that represents semantic relations (synonymy, hyponymy, hypernymy) in a set of words can be used for semantic matching. This kind of concept lattice is the result of the formal concept analysis technique used for determining semantic relations in a set of words extracted from a monolingual dictionary. It is shown how relations between concepts can be mapped into semantic matching relations (equivalence, disjointness, more specific, less specific). The results of using semantic matching with concept lattice in a spoken dialog manager system are shown. The models are represented in F-logic language and implemented in FLORA-2 system.

Keywords: Concept Lattice, Semantic Matching, Synonym Extraction, F-logic, Natural Language Dialog System

1 Introduction

This paper describes how formal concept analysis can be applied in the domain of natural language processing. An approach of semantic analysis using concept lattice as background knowledge is proposed. An important problem of semantic analysis is semantic heterogeneity that includes managing the diversity in knowledge. Therefore a process of semantic matching is defined. Semantic matching in this paper denotes an operation of matching two lexical units that have equal or similar meaning. Semantic matching may have a more general definition, such as matching operation that takes two graph-like structures (e.g., conceptual hierarchies, database schemas or ontologies) and produces mappings among the nodes of two graphs that correspond semantically to each other [8]. However, the main idea presented in this paper is one of semantic matching as an operation that matches different concepts that are semantically close.

Formal concept analysis is a technique that includes lattices and order theory as a tool for data analysis. It is based on sound mathematical theory, introduced to information science by Ganter and Wille [6,7]. The idea of using FCA in the domain of natural language processing has been already discussed in [5,3]. Further, the idea of using concept lattices for semantic relations capturing and in linguistic applications is presented in [12,14,18,19,20,21,17]. In [21] it is described how WordNet can be formalised using FCA. A similar approach has been used in [14] for capturing semantic relations between words given in a Croatian monolingual dictionary using formal concept analysis in FLORA-2 system. We presented how semantic relations that are automatically extracted from a dictionary can be formalised and visualised using concept lattice. These results can be further used for semantic analysis in the dialog system. In this work, a deductive object-oriented logic programming language named F-logic, for semantic analysis is used. F-logic [11] provides a natural way of defining a conceptual model of data semantics and Web data manipulation. Further, F-logic is a formalism that can capture formal concept analysis. All rules for semantic analysis are defined in F-logic and implemented in FLORA-2 system. It is shown how proposed semantic analysis with a concept lattice can be used as a part of the spoken dialog system for weather information in Croatia.

The second section of this paper introduces semantic analysis and semantic matching process. The third section presents how semantic relations defined in the monolingual dictionary can be represented using concept lattice. Section four presents how concept lattice can be implemented as a part of a spoken dialog system for the Croatian language. Finally, some possible improvements are discussed and some future work plans are presented.

2 Semantic analysis and semantic matching

Semantic analysis can be viewed as a task of translating a natural language sentences into a formal meaning representation language [1]. An important issue of semantic analysis and natural language understanding in general is how to treat semantically close words or phrases. Therefore a natural language understanding module needs to have additional knowledge about semantic relations between words. In this paper, an elementary process of linking semantically close words (concepts) is called semantic matching.

For two words w_1 and w_2 there are three possible relations that describe cases of semantic closeness: equivalence ($\equiv$), more specific ($\sqsubseteq$), partial overlapping ($\sqcap$). If two words have no semantic closeness then they are disjoint ($\perp$). Additionally, a relation less specific ($\sqsupseteq$) can be introduced as an inverse relation of more specific. For every two words w_1 and w_2 there is only one possible relation and it depends on how close these words are in meaning. Therefore, it is possible to define a function that connects two words by assigning a semantic relation that holds between them.

Definition 1. *Let W be the set of all possible words and let R be the set of relations, $R = \{\equiv, \sqsubseteq, \sqsupseteq, \sqcap, \perp\}$. Mapping f_{rel} from $W \times W$* to *R assignes to each pair of words $(w_1, w_2) \in W \times W$ an approprate relation from R that holds between* w_1 and w_2.

There are also measures defined for measuring semantic distance [16]. This is not considered in this paper, but eventually function f_{rel} can be extended with a measure of closeness and that is a plan for future work. Furthermore, semantic matching may be described with semantic correspondences called mappings attached to one of the following semantic relations: disjointness, equivalence, more specific, less specific, overlapping [8]. Although semantic matching has a broader definition as described in the introduction, the basic task of semantic matching in the context of semantic analysis is to connect the word (phrase) with a set of words (phrases) that are semantically close to it. For that purpose another mapping is defined.

Definition 2. *Let W and R be the sets as in the previous definition. Mapping f_{match} from $W \times R$ to a partitive set of W, $\mathcal{P}(W)$ assigns to each pair (w, r) where $w \in W$ and $r \in R$ a set of words W_m that for each word w_m from W_m holds $f_{rel}(w, w_m) = r$.*

Using mapping f_{match} it is possible to link each word with a set of words that are semantically close into it. Therefore, mapping f_{match} may be used to accomplish an operation of semantic matching. In the next section it is shown how all relations between words, that are a result of semantic matching operation, can be presented and visualised using concept lattice. Furthermore, a concept lattice that describes semantic relations can be used for semantic analysis.

3 Monolingual dictionary formalization using formal concept analysis

3.1 From a monolingual dictionary to a concept lattice

In this section an approach for dictionary formalization using the formal concept analysis (FCA) technique is presented. A similar approach may be defined by using WordNet as a resource of semantically close words and their relationships. There are many research projects that deal with WordNet and some of them use formal concept analysis as a technique for exploiting semantic relations, visualization and other research[9,12,21]. In this research a monolingual dictionary is used instead of WordNet. One reason is that WordNet for Croatian language is not completely finished and available. Eventually, there is more information that can be extracted from a dictionary than from WordNet. Moreover, the described approach can be used to define or update Croatian WordNet because FCA based semantic matching gives synsets as the final result.

Data in dictionaries are usually presented with implicitly defined structure. Important attributes of a word are organized following implicit dictionary structure. Each word may have more than one meaning. Each meaning has its own description with its own set of attributes described. This complex way of representing a word defines a dictionary structure that can be captured using formal grammar. A monolingual Croatian dictionary [2] is used for automatic extraction of semantically close words. The formal grammar is defined in F-logic in order to capture this structure. After formalizing the structure of a dictionary, the final goal was to extract words that are semantically close.

Set of semantically close words are further analyzed using FCA technique. Different semantic relations can be found in dictionaries (synonyms, near synonyms, hypernym, hyponym, etc.). These relations form a hierarchical structure between semantically close words. In [14] it is shown how that can be modeled as a formal context and viewed as a concept lattice.

If a word in a dictionary has more than one meaning and if any of these meaning is descriptive (not explained with another synonym), than a set of special marks is introduced (as *zd*, *zd*1, *zd*2,...). These marks are appearing in the formal context and concept lattice, denoting that there are more meanings for some word but with no synonyms in a dictionary.

Figure 1 shows three possible relations between two words (w_1, w_2) that have semantic overlapping. In the first case (a) word w_1 can replace word w_2 in any context and vice versa. In the second case (b) word w_2 can be replaced with word w_1 in any context, but not vice versa since word w_1 has a more general meaning. In the third case (c) both words w_1 and w_2 have some additional meaning and therefore can be replaced with each other only in certain contexts. These cases actually reflect the relationships between two synonyms or hyperonym and hyponym described as relations of equivalence, less specific and more specific respectively .

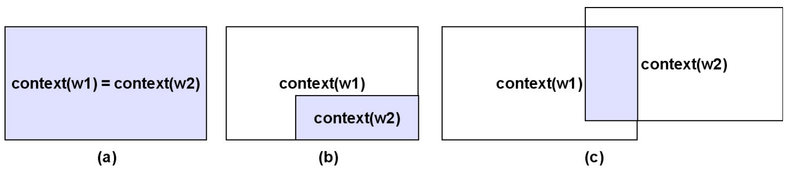

Fig. 1. Cases of semantic overlapping

The idea was to define formal context in a way that concept lattice as a result shows a naturally established hierarchy in a set of words. Firstly, the general case is analysed in order to present the basic idea of the proposed approach. Let $W_n = w_1, w_2, ..., w_n$ be a set of n words, then a formal context $C = (O, M, I)$ may be defined for set W_n. Set O is a set of words from W_n, $(O = W_n)$ and set M is a set of words that overlap with words from W_n. In some generalized cases set of attributes can be a subset to a set of objects, but here it is $M = O$. Rules for transforming dictionary data into formal concepts are defined in F-logic as it is shown in [14]. The transformation process assumes adding relations of reflexivity, symmetry and eventual transitivity. A final model should also capture a possibility of a different representation of meanings. Apart from this, incompleteness and irregularities can also appear in a dictionary and thus, have to be included into a model. Therefore, many additional rules for handling these specific situations are formed.

The final result of FCA technique applied to a dictionary is a concept lattice of words. The described model of a concept lattice reflects semantic closeness between words, therefore, it may be called a semantic relation concept lattice. On the higher levels of the lattice there are words with more general meaning and on the lower levels there are words with more specific meaning. These relations naturally correspond to relations defined within the semantic matching operator. The interpretation of a concept lattice is given in the next section.

3.2 Concept lattice interpretation

For the purpose of this paper, a small set of words is translated into English. A set of words with similar relations between them is chosen in order to show the main idea of semantic closeness representation using concept lattice. However, there may be slight differences in lattice interpretation using English language from which we have using Croatian language in the original example. An example of formal context is described for a given set of words $W_1 = \{infinite$, *endless*, *prominent*, *noted*, *enormous*, *strong*, *huge*, $well - known$, *eminent*, *big*, *high*, $remarkable\}$. All these words translated into Croatian language have semantic similarity detected in the Croatian monolingual dictionary. This particular set of words has been chosen as an example because it generates a rich concept lattice structure. Other sets of words with semantic similarity from Croatian monolingual dictionary contains a smaller number of words. Some smaller set of words are presented in the next section where an application of concapt lattice-based semantic analysis for the weather forecast domain is shown.

Using the rules for defining concepts shortly presented in a previous section, a set of 19 concepts is generated. Relations between concepts defined by a conceptual lattice are shown in Fig. 2.

Using the proposed technique, it is possible to define a concept lattice for any set of words that have semantic similarities. Each concept connects words from a dictionary in a way that concept extent corresponds to the words given in an intent set. For example, concept $k14$ with extent defined as $k14[extent \rightarrow [infinite, endless]]$ links words $infinite$ and $endless$ in a way that these two words have semantic overlapping with the same set of words given as the intent of the same concept, $k14[intent \rightarrow [infinite, endless, enormous, huge, big, zd1]]$. Moreover, these are the only two words that have semantic overlapping with this exact set of words represented as the concept intent. Hierarchical relationships defined between concepts in the concept lattice reflect the possible relations between words that can be attached using function f_{rel} defined in the second section. The hierarchy is defined in the way that the words belonging to extents of concepts of a higher level are less specific ($\sqsupseteq$) than the words which belong to lower level concepts. For example, in Croatian language the word *big* that is an extent of a concept $k10$ has less specific meaning than the word *huge* that belongs to an extent of concept $k2$, that is a subconcept of a concept $k10$. In further analysis of the concept $k10$ it can be noticed that an extent of the concept $k10$ is only the word *big*. Hierarchy of a lattice provides information that the word *big* semantically links all words from set W_1. For some words in a lattice the word *big* includes all their meaning and for other words it includes only partial meaning. For example, the word *big* can replace words *prominent*, *remarkable* in every context, but cannot replace the words *endless*, $infinite$ in every context. In the set of words that is given in this example there is no word that represents generalization of all the words because the concept $k1$ is an empty concept. The first lower level has five different concepts ($k14$, $k17$, $k10$, $k19$ and $k7$). These concepts can not be compared. The extents of these five concepts deal with the words that have a more general meaning.

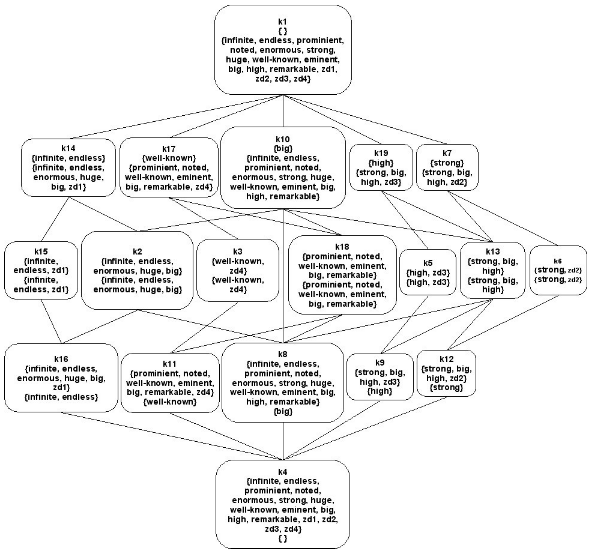

Fig. 2. A concept lattice for set of words W_1

Concept lattice can be presented in another way, as it is shown in Fig. 3. This kind of representation is called reduced lattice representation [4]. In the reduced lattice, each word is shown only once. The extent of a concept is formed by collecting all objects that can be reached by edges that connect that concept with concepts on a higher level. It is important that in reduced representation some relations between sets of words are visualized in a clearer way. Besides, this kind of representation shows only necessary words. In a reduced lattice, each word appears only once in the exact position that shows where it belongs in a word hierarchy. It is expressed that words *enormous* and *huge* are synonyms and that can be described as an equivalence relation. Set of words S={*prominent*, *noted*, *eminent*, *remarkable*} is also a set of synonyms with some exceptions. Words $well - known$ and *big* are more general than the words in set S. This means it is possible to replace every word from a set S with word $well - known$ or *big* in every context with no change in meaning. This reflects a relation more specific (less specific, on the contrary) between each word from W_1 and the word $well - known$. Henceforth the words $well - known$ and *big* semantically overlap, but have separate meaning nevertheless. It refers to an overlapping relation.

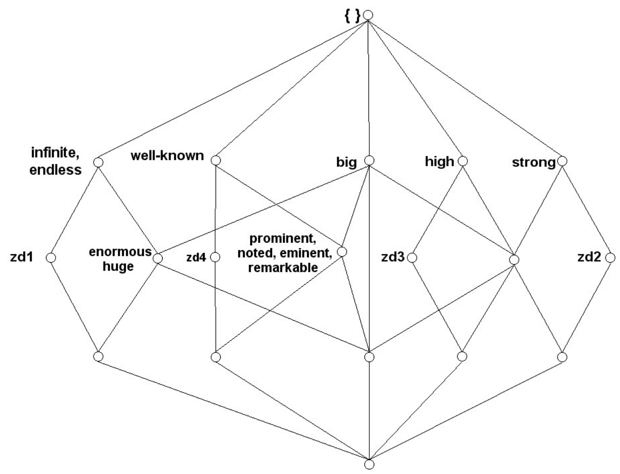

Fig. 3. A reduced concept lattice for set of words W_1

4 An application of semantic relation concept lattice

One possible application of semantic relation concept lattice is to use it as background knowledge in the process of semantic analysis. Semantic analysis is a fundamental part of the natural language understanding module. In this section an example of semantic matching using semantic relation concept lattice in the Croatian language dialog system for the weather forecast domain is described. The current status of a Croatian spoken dialog system prototype is presented in [13].

There are different approaches defined for semantic analysis that can be implemented in the natural language understanding module [10]: syntax-driven semantic analysis, semantic analysis based on formal grammar and information extraction. The Croatian weather data semantic analysis combines information extraction slot filling technique with grammar [15]. This combined approach is chosen mainly because of the limited weather domain and highly flective nature of the Croatian language. Information extraction is used with limited domain and when no detailed comprehention is needed. In information extraction process, knowledge can be described with simple templates. Templates consist of frames with slots that need to be filled with data from the text. In those situations only relevant information from the input text is used for filling the slots and the rest of the text is ignored. Information extraction with the slot filling technique is used in many semantic parsers of spoken dialog systems [22].

The slot filling technique is focused on predefined keywords and matching tempaltes. The problem of such a key word-based matching interpretation are words with semantic similarities. In the proposed approach knowledge of semantic similarities is automatically extracted from a dictionary and stored in the concept lattice. Therefore, for each word w it is possible to use f_{match} function to get all words that have semantic similarities (descibed relations from previously introduced set R) with the word w. One simple example of using f_{match} function to interpret the question from the weather forecast dialog system is shown in Table 3. Information about wind and weather is given as a small part of the weather forecast for the Adriatic coast. Wind names are specific for the Adriatic coast and therefore are not translated into English (note that *jugo* and *široko* are synonyms for the same wind and *bura* is a wind name, also). The result of applying concept lattice for matching words from the question, more precise answers are obtained.

Table 1. An example of question answering with and without using concept lattice (CL)

Example of question	Answer with no CL	f_{match} mapping from the CL	Answer using CL
An example of weather forecast data that is used for answering questions: *Jugo* is going to blow today with occasional precipitation on the Adriatic coast. Tomorrow, *bura* is a possibility.			
Does *široko* blow on the Adriatic coast?	No	$f_{match}(široko, \equiv) = jugo$	Yes
Is there precipitation on the Adriatic coast?	No	$f_{match}(rain, \sqsubseteq) = precipitation$	Yes
Is wind going to blow tomorrow?	No	$f_{match}(wind, \sqsupseteq) = bura$	Yes

5 Conclusion

In this paper an approach of concept lattice-based semantic matching is introduced. The main motivation of this research was to improve the process of natural language understanding in the previously developed Croatian language dialog system for the weather forecast domain. At first, the idea of how to use semantic match operators in the process of semantic analysis is introduced. Secondly, the idea of using FCA technique for representing semantic relationships between words in a dictionary is presented. Three different models of semantic relationships are modeled as three different formal contexts using F-logic. This way, a set of words can be represented using formal context designed in order to represent semantic hierarchy between words. The final result is a concept lattice that shows a semantic overlapping between words. Implementation and results are presented in the fourth section. The conceptual model is represented in F-logic and after that implemented in FLORA-2 system.

The final result is more precise semantic analysis and more precise answer generation with concept lattice for semantic matching (Table 1). This hypothesis has been proved for a set of examples, but no evaluation for the whole system has been done yet. The evaluation of concept lattice-based semantic matching is a topic for further research.

References

1. Allen, J.: Natural language understanding. Benjamin/Cummings series in computer science, Benjamin/Cummings Pub. Co. (1995), http://books.google.hr/books?id=l4lQAAAAMAAJ
2. Anić, V.: Rječnik hrvatskoga jezika. Novi Liber (2005)
3. Boutari, A.M., Carpineto, C., Nicolussi, R.: Evaluating term concept association measures for short text expansion: two case studies of classification and clustering. In: Proceedings of the Seventh International Conference on Concept Lattices and their Applications (CLA 2010) (2010)
4. Carpineto, C., Romano, G.: Concept data analysis: theory and applications. Wiley (2004), http://books.google.hr/books?id=-F8OoVXQioAC
5. Falk, I., Gardent, C.: Combining formal concept analysis and translation to assign frames and thematic grids to french verbs. In: Amedeo Napoli, V.V. (ed.) International Conference on Concept Lattices and Their Applications CLA 2011. pp. 223–228. INRIA Nancy - Grand Est and LORIA (2011)
6. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer (1999), http://books.google.hr/books?id=cN1QAAAAMAAJ
7. Ganter, B., Wille, R.: Applied lattice theory: Formal concept analysis. In: G. Gatzer editor, B. (ed.) General Lattice Theory (1997)
8. Giunchiglia, F., Shvaiko, P., Yatskevich, M., Giunchiglia, F., Shvaiko, P., Yatskevich, M.: S-match: an algorithm and an implementation of semantic matching. In: In Proceedings of ESWS. pp. 61–75 (2004)
9. Hotho, A., Staab, S., Stumme, G.: Explaining text clustering results using semantic structures. In: In Principles of Data Mining and Knowledge Discovery, 7th European Conference, PKDD 2003. pp. 217–228. Springer-Verlag (2003)
10. Jurafsky, D., Martin, J.: Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall series in artificial intelligence, Pearson Prentice Hall (2009), http://books.google.hr/books?id=fZmj5UNK8AQC
11. Kifer, M., Lausen, G., Wu, J.: Logical foundations of object-oriented and frame-based languages. Journal of the ACM 42, 741–843 (1995)
12. Martin, B., Eklund, P.: Applying formal concept analysis to semantic file systems leveraging wordnet. In: Proceedings of the 10th Australasian Document Computing Symposium (2005)
13. Meštrović, A., Bernić, L., Pobar, M., Martinčić-Ipšić, S., Ipšić, I.: A croatian weather domain spoken dialog system prototype. CIT. Journal of computing and information technology 18, 309–316 (2010)
14. Meštrović, A., Čubrilo, M.: Monolingual dictionary semantic capturing using concept lattice. International Review on Computers and Software (I.RE.CO.S.) 6, 173–184 (2011)
15. Meštrović, A., Martinčić-Ipšić, S., Čubrilo, M.: Weather forecast data semantic analysis in f-logic. Journal of Information and Organizational Sciences 31, 115–129 (2007)

16. Mohammad, S., Gurevych, I., Hirst, G., Zesch, T.: Cross-lingual distributional profiles of concepts for measuring semantic distance. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007)
17. Old, J.: Homograph disambiguation using formal concept analysis. In: Missaoui, R., Schmid, J. (eds.) 4th International Conference on Formal Concept Analysis, Lecture Notes in Computer Science. vol. 3874, pp. 221–232. Springer-Verlag (2006)
18. Potemkin, S.: Concept lattice implementation in semantic structuring of adjectives. In: Ignatov, D., Kuznetsov, S., Poelmans, J. (eds.) Concept Discovery in Unstructured Data. pp. 63–70 (2011)
19. Priss, U.: Linguistic applications of formal concept analysis. In: Wille, G.S. (ed.) Formal Concept Analysis, Foundations and Applications. pp. 149–160. Springer Verlag (2005)
20. Priss, U., Old, L.J.: Modelling lexical databases with formal concept analysis. Journal of Universal Computer Science, Vol 10, 967–984 (2004)
21. Priss, U.E.: The formalization of wordnet by methods of relational concept analysis. In: WordNet: An Electronic Lexical Database and Some of its Applications. pp. 179–196. MIT Press (1998)
22. Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V.: Galaxy-ii: A reference architecture for conversational system development. In: in Proc. ICSLP. pp. 931–934 (1998)