Use of Wordnet for Retrieving Words from Their Meanings

İlknur Durgar El-Kahlout and Kemal Oflazer

Sabancı University
Faculty of Science and Nature
34956 Orhanlı-Tuzla İstanbul, Turkey
Email: ilknurdurgar@su.sabanciuniv.edu, oflazer@sabanciuniv.edu

Abstract. This paper presents a Meaning to Word System (MTW) for Turkish Language, that finds a set of words, closely matching the definition entered by the user. The approach of extracting words from "meaning"s is based on checking the similarity between the user's definition and each entry of the Turkish database without considering any semantics or grammatical information. Results on unseen user queries indicate that in 66% of the queries the correct responses were in the first 50 of the words returned, while for queries selected from the word definitions in a different dictionary in 92% of the queries correct responses were in the first 50 of the words returned. Our system make extensive uses of various linguistics resources including Turkish WordNet.

1 Introduction

Suppose one can not remember a word but knows a variety of contextual phrases that approximate his or her understanding of the word and wants to find the appropriate word (or words) that has similar meaning with his/her definition. For this problem, it will be of no use to attempt searching in a traditional dictionary to find the word. Traditional dictionaries are helpful for finding the meaning of a word but we need an application that works in the opposite direction.

Some examples of definitions taken from users and the corresponding meanings of those words taken from dictionary [1] are listed below.

- akımölçer (ammeter)
  - **User Definition:** *akımı ölçmek için kullanılan alet* (a device that is used to measure the current).
  - **Dictionary Definition:** *elektrik akımının şiddetini ölçmeye yarayan araç, amperölçer* (a device that measures the intensity of electrical current, amperemeter).
- istifa (resignation)
  - **User Definition:** *çalıştığı işten kendi isteğiyle ayrılmak* (leaving one's job voluntarily).
  - **Dictionary Definition:** *kendi isteğiyle görevden ayrılma* (leaving voluntarily, of a position).

The definitions collected from a set of users showed us that users usually define the words very similar to the actual dictionary definitions in terms of meaning. By using this similarity, we implement a system called Meaning to Word (MTW) for Turkish to find the appropriate words whose definitions match the given definition.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 118–123.

2 Meaning to Word

While finding the appropriate words, MTW deals with two challenging problems: (i) locating a number of candidate words whose definitions are "similar" to the definition in some sense, (ii) ranking these candidate words using a variety of ways to return a list sorted in terms of similarity. Our approach for extracting words from meanings is based on checking the similarity between the user definition and each entry of the dictionary by making a number analyses without taking into consideration the semantics or the context.

MTW works as follows: A user definition is given as an input to the system. The user definition is processed to construct a query. With this query, the database is searched and a list of candidate words is generated. The candidate words are sorted in terms of similarity and the list is returned to the user as a result. It should be noted that all the processing steps are fully automated, no human intervation or manual encoding is required. We use NLP techniques to enhance the effectiveness of term-based information retrieval.

2.1 Databases and Other Sources of Information

We use two resources in retrieving appropriate words for the user request. These sources are the explatonary Turkish Dictionary and Turkish WordNet.

MTW uses the Turkish dictionary to search in and match the corresponding meanings to the user's request. Dictionary has alphabetically sorted words and their meanings with 89,019 entries, 82,489 unique words and 21,653 unique stems.

Also, MTW uses Turkish WordNet to find the relations between words. Turkish WordNet [2] is structured in a similar way as the WordNet [3] around the notion of a synset. Synsets are linked across basic semantic relations such as hyponymy/hypernymy,antonymy and meronymy.

2.2 Query Generation

MTW does not use the user definition as it is; a set of useful information from the definition is selected with simple NLP techniques to form a query [4]. The steps are as follows:

Tokenization: We divide the symbols into two parts: Word symbols and non-word symbols. Characters other than letters and digits are treated as non-word symbols (e.g. punctuation marks)and eliminated from the definition because they are unnecessary for further retrieval.

Stemming: Because of the structure of Turkish, the words of the user's definition and the corresponding definition in the dictionary may have the same stem but different affixes. Stemming enables matching different morphological variants of the original definition's words.

Stop Word Removal: Stop words are words that contribute nothing or very little meaning; they should be removed from the query and dictionary definitions. If a word occurs frequently in a dictionary or has little meaning conceptually (such as prepositions, determiners), then it is not an informative word. The top $200 - 300$ frequent words in the dictionary and conceptually little meaning words are selected as stop words and removed from dictionary definitions and the query.

Stemming process takes place before the stop word removal because of the structure of Turkish. For example, the words *bir* (one), *biri* (one of them), *birileri* (some people) have the frequencies 19901, 12 and 2, respectively. Although all of the words have the same stem *bir* (one), it is possible to eliminate only the word *bir* (one) with the given frequencies.

2.3 Query Processing

While searching for the appropriately matching meaning, rarely all of the query words match the relevant meaning. For this reason, an approximate match is more suitable than the exact match of user's request with the dictionary meanings. In MTW, sub-queries are generated by using different combinations of words from the original query. Then, MTW sorts the sub-queries in order to their informativeness.

Subset Generation: MTW generates all $2^n - 1$ sub-queries for a n word query, where n is the number of words remaining in the query after stop word removal. Table 1 shows sub-queries generated from the query *yazlık büyük ev* (large house for summer). **Subset**

Table 1. Subset generation table for query *yazlık büyük ev* (large house for summer)

Subset number	yazlık	büyük	ev	Generated subset
1	1	1	1	yazlık büyük ev (large house for summer)
2	1	1	0	yazlık büyük (large for summer)
3	1	0	1	yazlık ev (house for summer)
4	1	0	0	yazlık (for summer)
5	0	1	1	büyük ev (large house)
6	0	1	0	büyük (large)
7	0	0	1	ev (house)

Sorting: Searching the meanings with an unordered sub-query list is not efficient as we can not estimate which sub-query can give the correct meaning. For this reason, we should start from the most informative sub-query first. The sub-queries are sorted in order to the number of words that they contain. This lets the system to find the meanings matching maximum number of words before the others. If there are two meanings that match the same number of words then the system decides which of the sub-query is more informative than the other.

Table 2. Frequencies of each word of the query *yazlık büyük ev* (large house for summer)

Word	Word Occurrence	Stem Occurrence
yazlık (for summer)	9	12
büyük (large)	931	1168
ev (house)	157	734

From Table 2, the word *yazlık* (for summer) is more informative than the words *büyük* (large) and *ev* (house), and the word *ev* (house) is more informative than the word *büyük* (large). For multi-word sub-queries, the logarithms of word frequencies are added and the result is used to define the information measure of the subset. We use the sum of word frequency logarithms as the frequencies of words are too small and directly multiplying the frequencies will cause information loss. The sorting formula is:

$$relevance\_of\_subset(j) = \frac{\sum_{i \in subset_j} \log(freq_i)}{N_j}. \quad (1)$$

where, $freq_i$ is the frequency of i^{th} word in dictionary and N_j is the number of words of the j^{th} subset.

2.4 Searching for 'Meaning'

Simplest idea for finding the similarity between two phrases is to match the common words of both, and return the best matching meaning. But, user's definition and actual meaning of a word generally have same concepts with different words [5]. For example:

- **User Definition:** *daha önce hiç evlenmemiş olan kişi* (a person who has never been married).
  - **Generated Query:** *daha, {önce, ön}, hiç, evlen, ol, kişi* (yet, {before, front}, never, marry, be, person).
- **Actual Definition:** *evlenmemiş kimse* (unmarried person).
  - **MTW Representation:** *evlen kimse* (marry, person).

At first sight, only the word *evlen* (marry) is matching with the actual meaning. But the words *kişi* (person) and *kimse* (someone) are similar words. Standart matching algorithm using only stems will fail to find this similarity. But for the efficiency of the retrieval, these words should also be counted as "matched". Use of Turkish WordNet helped us to find the possible matching words. In our method, we use the synonym words from the Turkish WordNet and expand the query. In Turkish WordNet there is a synset *{kişi (person), kimse (someone), şahıs, birey (individual), insan (human)}* containing both of the words. The original query contains only *kişi* (person) but the extended one will contain all the synset members including *kimse* (someone). The method is applied to all the words in the original query. The enhanced query will retrieve dictionary definitions with higher ranks.

2.5 Ranking

MTW uses three criteria to rank the candidate definitions: (i) the number of matched words is calculated. If any definition has more common words with the query than others, then this definition is more relevant; (ii) the length of the candidate definition is determined. If two candidates have the same number of matches with the user definition, the shorter candidate is ranked before the longer one; (iii) the longest common subsequence of the candidate definition and user definition is calculated. The definition that have longer common subsequence is ranked before the shorter ones.

Table 3. Results of MTW with all stems included

Rank	train_set	test_set	dict_train_set	dict_test_set
1 – 10	14 (28%)	24 (48%)	45 (90%)	41 (82%)
11 – 50	9 (18%)	9 (18%)	2 (4%)	5 (10%)
51 – 100	3 (6%)	3 (6%)	1 (2%)	2 (4%)
101 – 300	7 (14%)	2 (4%)	2 (4%)	1 (2%)
301 – 500	0 (0%)	1 (2%)	0 (0%)	1 (2%)
501 – 1000	4 (8%)	5 (10%)	0 (0%)	0 (0%)
over 1000	4 (8%)	1 (2%)	0 (0%)	0 (0%)
not found	9 (18%)	5 (10%)	0 (0%)	0 (0%)

3 Performance Evaluation

3.1 Setup

The experiments were carried out on two different test sets: test_set and dict_test_set. In addition, two train sets are used: train_set and dict_train_set. In the experiments 50 user definitions are used for each set. Queries for test_set and train_set are taken from real users. Users are given 50 different words and asked to define these words. Definitions for dict_test_set and dict_train_set are taken from a dictionary [2]. The dictionary definitions of the same 50 words that are given to the users are used as definitions.

3.2 Results

Sometimes stemming algorithms can produce different meaning stems for a word, such as for the query *en yüksek yer* (most highest place), the stemmer gives two different stems *yük* (load) and *yüksek* (high) for the word *yüksek*(high, if we are load) but only the word *yüksek* (high) is the correct stem. We test our method with two different approaches. In the first test, all of the stems returned from the stemmer (i.e., *yük* (load) and *yüksek*(high)) are included in the query. In the second test, a simple heuristic approach is used. We assume that the longest stem (i.e., *yüksek* (high)) returned from the stemmer is the correct stem and include only this stem to the query.

Tables 3 and 4 show the results of the method with all stems and only longest stems, respectively.

With our method, we can match the 66% of the user definitions and 92% of the dictionary definitions by using all stems, and 68% of the dictionary definitions and 90% of the dictionary definitions by using longest stem in the first 50 results. There is a decrease when we select the longest stem because the longest stem may not be the correct stem for every word. Although there is a little increase in the first 50 rank in the test_set, the performance decreases in the first 10 rank.

4 Conclusions

In this paper, we presented the design and implementation of a Meaning to Word system that locates a Turkish word that most closely matches the appropriate one, based on a

Table 4. Results of MTW with only longest stem included

Rank	train_set	test_set	dict_train_set	dict_test_set
1 – 10	15 (30%)	22 (44%)	45 (90%)	40(80%)
11 – 50	8 (16%)	12 (24%)	1 (2%)	5(10%)
51 – 100	3 (6%)	2 (4%)	0 (0%)	4(8%)
101 – 300	6 (12%)	2 (4%)	3 (6%)	1(2%)
301 – 500	1 (2%)	2 (4%)	0 (0%)	0(0%)
501 – 1000	2 (4%)	3 (6%)	0 (0%)	0(0%)
over 1000	6 (12%)	2 (4%)	0 (0%)	0(0%)
not found	9 (18%)	5 (10%)	0 (0%)	0(0%)

definition entered by the user. Using only simple and symbolic methods, the performance results of MTW on unseen data from real users are rather satisfactory. The results on unseen queries from a different dictionary shows that the methods used while implementing MTW are reasonable. MTW has the advantage of free stemming and expansion that gives a great flexibility to retrieval. By stemming and query expansion in MTW, the user's definition can match the correct word(s) even if the terms of the dictionary definition does not contain the same words with same affixes. Besides MTW has the disadvantage of false matches. Because of the noise from the wrong stems and unrelevant synonyms, MTW can produce many irrelevant candidates. MTW works best if the request is typed similar to the actual definition. MTW can be used in various application areas such as computer-assisted language learning, crossword puzzle solving, or as a reverse dictionary.

References

1. Püsküllüoğlu, A.: Arkadaş Türkçe Sözlük. Arkadaş Yayınevi (1995).
2. Turkish WordNet, [online]: `http://fens.sabanciuniv.edu/TL`. (Accessed: September 5, 2003).
3. Miller, G.: WordNet: A Lexical Database for English. Communications of the ACM 38(11) pp 39–41(1995).
4. Strzalkowski, T.: Natural language information retrieval. Information Processing & and Management 31(3), 397–417 (1995).
5. Voorhees, E. M.: Using WordNet for Text Retrieval. In: Fellbaum, C. (ed.): WordNet – An Electronic and Lexical Database, MIT Press. Cambridge, Mass., (1998), pp 285–303.