Substitutions Acquisition Method for an Informational Retrieval System

Artem Churkin and George Mazurkevich

Dept. of Computational Linguistics,
Mail.Ru Group,
Leningradskiy prospect 47, Moscow, Russia
{a.churkin,g.mazurkevich}@corp.mail.ru

Abstract. The paper is devoted to automatical acquisition of substitutions for an information retrieval system. One method, based on the distributional hypothesis and some additional proposals about source data, is suggested. Method is focused on using nonstructured texts as resources, especially logs of users' queries. A few practical results of using the method are demonstrated. Strength and weakness of the method are discussed.

Keywords: query expansion, substitutions acquisition, contextual language model, context distance, similarity measure, information retrieval, natural language processing

1 Introduction

Modern web-based information retrieval (IR) systems are complex objects intended to satisfy information needs of wide audience of users. Such systems contain different components and modules which perform different actions (web-crawling, indexing, parsing queries, retrieving documents, ranking, etc.) and interact with each other. When the audience becomes wider such system starts facing more various and complex information needs. Nevertheless most modern web-based information retrieval systems have simple interfaces, which contain a text field and one or two buttons. In spite of the limitations, search quality derived by the systems is suitable enough for most of their users. High quality level is achieved by developing more and more sophisticated components and modules mentioned above. One of these components is "query parser", which is intended to "understand" a query received from a user. Query understanding is a complex concept and combines different meanings, one of them is "to extend a query with substitutions". Main point of extending a query is demonstrated in the following example.

Example 1. Query–document relevance.
Query: "popular guesthouses in spb"
Document title: "Saint-Petersburg Hotels"
Related words in the document's content: "Saint-Petersburg, hostel, price, etc."
Extended query: "'popular (guesthouses|hostels) in (spb|Saint-Petersburg)"

According to the title and content user can assume that the document in the example 1 is relevant to the query and deserves to be shown in search results. But the relevance is implicit and IR system can't detect it without additional information that words 'guetshouse' and 'hostel' mean nearly the same. One of the approaches to the problem is to extend the query by substitutions. The bottleneck of the extension procedure is to build and maintain in actual state a base of substitutions. To succeed in practical implementation of the approach one should solve this problem. One practical method for building and refreshing the base of substitutions is suggested in the paper.

Worth to point out is that query expansion is a proven and effective technique to overcome lexical gap between documents and queries. There are many works addressed to problems related to query expansion and particularly substitutions acquisition. Most of existing approaches are based on the key idea of mapping queries' terms to documents' terms. The work [1] should be mentioned as one of the implementations of the idea. To use such approaches in practice a training set of marked up pairs of queries and relevant documents is needed. There are some other approaches which don't require such structured information for practical using. And the method proposed in this work belongs to such type of the approaches. It is based on the idea common to the one described in [2], that similar words have identical contexts. The major difference is in the constructed model for mining substitutions.

The remainder of the paper is organised as follows. Section 2 addressed to the concept of substitution in an IR system. In the section 3 our method for automatical substitutions acquisition is constructed. Practical results and further work are discussed in sections 4 and 5.

2 Substitutions

The term substitution can be considered as synonym, but actually it is wider. It may be a pair of some real synonyms, a pair of words with different parts of speech, transliteration pairs (for languages using non-latin alphabet), different spelling variations of borrowed words which don't have a long-held spelling, abbreviation and its full form. Several examples in Russian (our focus in IR system) are given below. Some semantical types:

- Transliteration: ютуб - youtube
- Localization and translation: евровидение - eurovision
- Abbreviations: оон - организация объединенных наций
- Synonyms: список литературы - библиография
- Grammatical substitutions: гороховый суп - суп из гороха
- Joins/Splits: бибиси - би би си
- Spelling variations: хэтчбек - хетчбек

Semantic classification is usefull in substitutions base construction and its actualization procedure. It helps to specify sources of substitutions: highly structured linguistics bases (dictionaries), semistructured text resources (web-resources:

wikipedia.org, imdb.com, etc.), nonstructured texts (logs of queries from users), grammatical generation algorithms, etc. Also substitutions can be classified by replacement types:

- Unidirectional – replacement can be made only from x to y not reverse.
- Bidirectioanl – replacemant can be made in two directions from x to y and back.

Replacement classification is important in query expansion procedure.

3 Acquisition

When some explicit structure of a web-resource can be captured, extraction algorithms can use it, and statistical methods often become unnecessary or redundant. But in case of nonstructured texts statistics become a key instrument for acquisition. Logs of users' queries (daily, weekly, monthly) are an example of suchlike resources. It is one of the most appealing materials for following reasons:

1. amount of queries is always sufficient for using statisitcal methods;
2. it contains linguistic behavior information (special word use cases, typical errors, etc.);
3. they are always actual;

Threrfore suggested method for substitutions' acquisition focuses on using logs of queries as a source. The method can be used as a part of one of the two technics: candidates verification or searching bests for core element.

3.1 Empirical and Theoretical Foundations

Our method for automatical substitutions' acquisition is based on several proposals and hypothethis. At first it is the distributional hypothesis (Harris, 1954) described in works [3] and [4]. It can be formulated as follows.

Proposition 1. *Words that occur in the same contexts tend to be similar. According to substitution acqisition problem the term "similar words" (from previous works) is considered in our research as a set of words that have a good replacement ability.*

We need two following propsals for using statistical methods and for one logical transition in further reasoning.

Proposition 2. *Users of an IR system generate enough queries with similar meanings in defferent forms.*

Proposition 3. *A word frequency and a frequency of the word determinative contexts are correlated.*

A determinative context is the typical context the word is used in. The more often a word is used, the more it is used in its determinative context and vice versa. So it is one of the most probable contexts for the word.

Definition 1. *A context of lenght n is an ordered set (or vector) of n words. The wildcard symbol* $(*)$ *denotes any word.*

For example $(*, w, *)$ denotes any context of length 3, where word w is on second position.

Definition 2. *A set of contexts of lenght n, where only one word w is specified and placed on the position p, while other items are wildcards, is named as set of w's contexts on p'th position. When p is not specified then any position is allowed.*

According to the distributional hypothesis a set of w_1's contexts and a set of w_2's contexts (where w_1 and w_2 are similar words) must have significant common part at least for one of the sets. To formalize proposition 3 let's denote number of times when a context occurred in a corpus (breifly context's count) as $C(w_{left}, w_{mid}, w_{right})$ where $(w_{left}, w_{mid}, w_{right})$ is an argument. Then consider maximum likelyhood estimation (MLE) of probability for occurring the context (w_{left}, w, w_{right}) on condition of $(w_{left}, *, w_{right})$

$$P_{MLE}(w_{left}, w, w_{right} | w_{left}, *, w_{right}) = \frac{C(w_{left}, w, w_{right})}{C(w_{left}, *, w_{right})}. \quad (1)$$

Definition 3. *A context with only one wildcard, for example* $(w_{left}, *, w_{right})$*, which gives* $P_{MLE} \neq 0$*, when the wildcard is replaced by a word 'w', is named as allowable for 'w' context.*

So a determinative context for a word w is an allowable context, which gives one of the highest P_{MLE} value. The proposition 3 states that a count of w's context and a count of corresponding determinative context for w are correlated. Summarizing this with propositions 2 and 1 note that if words w_1 and w_2 are similar they have raletively large amount of common determinative contexts which P_{MLE} values are correlated. This is the base idea for building substitution extraction method. Next step is to organize source data in such a way that allows quick extraction of word's contexts and corresponding determinative contexts. Contextual language model allows to achieve those purposes.

3.2 Contextual Language Model

A contextual language model (CLM) is a specially organized n-gram language model that provides quick context search and estimation. Key concepts of a CLM:

- an n-gram is considered as a point of an n-dimensional space;
- a coordinate axis of the space corresponds to an index of a word in the n-gram;

– words are points of the coordinate axis.

So a CLM can be interpreted as an n-dimensional cube as demonstrated in Figure 1. Each cell of the cube keeps any estimated features corresponding to an

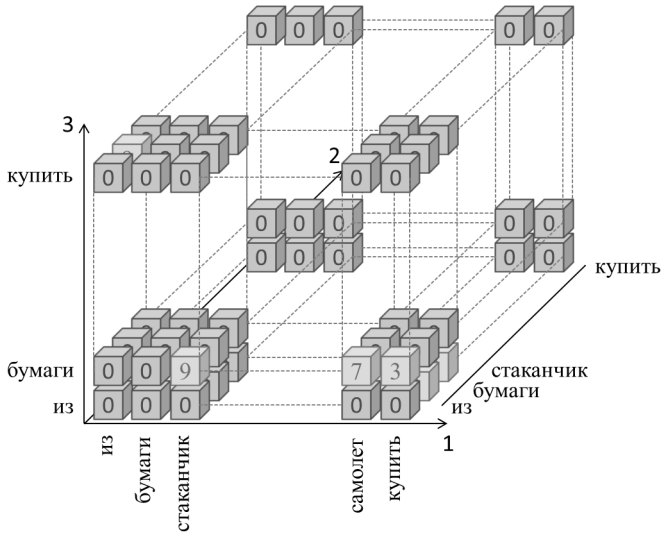

Fig. 1: Representation of CLM as 3d cube of 3-grams.

n-gram (simply count for example). Such data structure allows to get any context or any set of contexts as defined in section 3.1 with constatnt complexity $O(1)$. Therefore it is convinient for our calculation purposes. For example to obtain a set of contexts of type $(*, w, *)$ in 3d space, we only need to fix the word w on the axis 2 and get a secant hyperplane parallel to the axis 1 and 3. Taking into account the fact that most of cells of such cube have zero values, many optimization technics become available to keep a CLM in memory.

3.3 Substitution Metrics

Let's continue constructing the method for substitutions acquisition by defining similarity metrics. There are different works about similarity metrics based on words co-occurrence. For example, the work [5] contains wide overwiew of similarity metrics, development history, theoretical bases and practical aprocaches. All metrics considered in the work are based on the information theory approach. They use the notion of mutual information between words and contexts to measure similarity. This information approach is developed in further works. In the works [6] and [7] more sophisticated metrics are suggested and analyzed.

To build similarity metrics for substitutions acquisition we refused from the information theory approach in favor of mathematical statistics and the notion of correlation. Let's start with representation of a word w as a set of w's contexts. Consider a set of w_1's contexts: $\left(ctx_1^1, \ldots, ctx_L^1\right)$ and a set of w_2's contexts: $\left(ctx_1^2, \ldots, ctx_M^2\right)$, where ctx_j^i is j's allowable for w_i context. Next step is to intersect above sets and retrieve common contexts: $(ctx_1, \ldots, ctx_N)$. For each common context ctx_i we can obtain P_{MLE} estimations according to the formula 1 and decide is it determinative or not. Let N_d be a number of determinative contexts. So there are two samples of P_{MLE} values for the words w_1 and w_2, and we can estimate correlation between them. Since we work with the ordinal variables, we consider two nonparametric correlation formulas: Spearman's coefficient and Kendall's one. Let $\left(c_1^1, \ldots, c_{N_d}^1\right)$ be a set for w_1 and $\left(c_1^2, \ldots, c_{N_d}^2\right)$ be a set for w_2. Then Spearman's correlation is:

$$\rho = 1 - \frac{6}{N_d(N_d-1)(N_d+1)} \sum_{i=1}^{N_d} \left(R_i^1 - R_i^2\right)^2, \tag{2}$$

where R_i^1 is a rank of c_i^1 value in the sample for w_1, R_i^2 is a rank of c_i^2 value in the sample for w_2, N_d is a sample size. Kendall's correlation is:

$$\tau = 1 - \frac{4}{N_d(N_d-1)} \sum_{i=1}^{N_d-1} \sum_{j=i+1}^{N_d} \left[\left[c_i^1 < c_j^1\right] \neq \left[c_i^2 < c_j^2\right]\right]. \tag{3}$$

Both formulas can be used to calculate a distance (similarity metric) between words by the following formula:

$$sim\left(w_1, w_2\right) = \left(1 - \max\left(\frac{N_d}{L}, \frac{N_d}{M}\right)\right)\left(1 - Cc\left(w_1, w_2\right)\right), \tag{4}$$

where $Cc\left(w_1, w_2\right)$ is one of the variants of correlation coefficients. The formula 4 reflects the nature of a natural language and users behavior and allows to measure replacement ability of a word by another word in a query.

4 Results

Based on CLM and similarity metrics described in the section 3.3, we build a system for an automatical susbstitutions extraction. We use a corpus of users' queries with following characteristics:

- over 180 million of strings;
- each string is longer then 5 words;
- near 14 million of uniq word-forms.

The 3-dimensional CLM is constructed from the corpus, it contains statistic data of over 1 billion of 3-grams. The system allows to verify sets of substitution candidates. Also it allows to search nearest words for a given one. Some details

какой * сварить
* с тефтельками
грибной * рецепт
куринный * рецепт
вкусный куриный *
* из шпината
* с плавленным
* сжигающий жир
...

Table 1: Determinative contexts for “суп” and “супчик”.

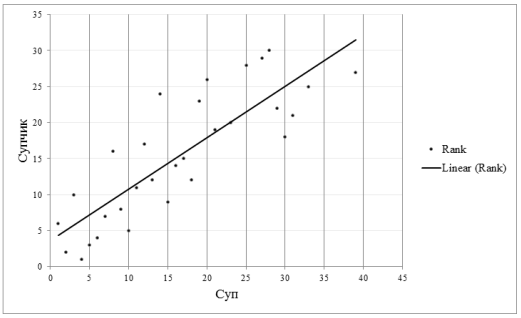

Fig. 2: Correlation of ranks of the contexts.

рукавицы

	варежки	перчатки	носки	вещи	шарф
$(*, w_{mid}, w_{right})$	0.2313	0.2601	0.2633	0.2643	0.2675
$(w_{left}, *, w_{right})$	0.2205	0.2327	0.2634	0.2598	0.2736
$(w_{left}, w_{mid}, *)$	0.2213	0.2479	0.2577	0.2752	0.2812
Total	0.2347	0.2515	0.2605	0.2658	0.2756

котик

	котенок	кот	щенок	малыш	мальчик
$(*, w_{mid}, w_{right})$	0.1436	0.1873	0.2005	0.1982	0.2264
$(w_{left}, *, w_{right})$	0.1470	0.2017	0.2104	0.2098	0.2229
$(w_{left}, w_{mid}, *)$	0.1501	0.1981	0.1962	0.2080	0.2166
Total	0.1447	0.1966	0.1971	0.2046	0.2220

супчик

	супик	суп	салат	суп-пюре	бульон
$(*, w_{mid}, w_{right})$	0.1853	0.2800	0.3147	0.3184	0.3269
$(w_{left}, *, w_{right})$	0.1888	0.2833	0.3225	0.3236	0.3317
$(w_{left}, w_{mid}, *)$	0.1729	0.2912	0.3248	0.3282	0.3244
Total	0.1831	0.2865	0.3216	0.3230	0.3271

Table 2: Top 5 nearest words for the given ones with metrics values based on Spirman’s correlation. First column describes type of a context.

of calculating similarity distance are given in Table 1 and Figure 2. The words “суп” and “супчик” are considered as the example. Table 2 demonstrates a few sets of nearest words according to the similarity metric.

Dealing with IR systems one of the most difficult points is checking the influence of done work on the search quality. We manage to deal with that by asking different people of different age and sex to estimate search queries. Those people get a set of queries in which the query tree has changed because of the particular work we have done. They can see two different versions of search output and decide according to their own "search preference" which version is

better. As we see that the search quality doesn't drop according to the assessions we can be sure to release the modified component.

5 Discussion

Our practical experience confirms well performance of the substitutions acquisition system based on the suggested method. Nevertheless to achieve good quality the system needs to be tuned. The common contexts classification threshold and the substitution acceptance threshold are the main parameters that require adjustment. The first one is needed for dividing common contexts into two sets: determinative and not, the second – for discarding bad substitutions' candidates. The adjustment procedure depends on characteristics of the source corpus and generally is the manual-based process.

Also there are some points for the essential improvement of suggested substitutions acquisition method. Now we work on several problems listed below:

1. The system is able to work only with unigrams. If we want to compare bigrams, trigrams etc we need to build bigger context language models. There is also a necessity of comparing n-grams of different order.
2. Contemporary CLMs are not able to use morphology analysis while comparing the words. Thus the russian words of different genders won't be able to be considered as simillar because of wrong context interpretation. Compare: "лучшая трехзвездочная гостиница" and "лучший трехзвездочный отель".
3. Not all the words of all parts of speech are compared well by this method. Adjectives may appear in simillar contexts but in fact they tend to have opposite meanings. Compare: "купить новый ford fusion" and "купить подержанный ford fusion".

References

1. Hang Cui, Ji-rong Wen, Jian-yun Nie and Wei-ying Ma: Probabilistic Query Expansion Using Query Logs. In Proc. of the Eleventh World Wide Web, pp 325–332 (2002)
2. Van Dang, Xiaobing Xue and W. Bruce Croft: Context-based Quasi-Synonym Extraction. CIIR Technical Report (2009)
3. Harris, Zellig S.: Distributional Structure. Word, vol. 10, pp. 146–162 (1954)
4. Sahlgren, Magnus: The Distributional Hypothesis. Rivista di Linguistica 20, pp. 33–53 (2008)
5. Ido Dagan, Shaul Marcus and Shaul Markovitch: Contextual Word Similarity and Estimation from Sparse Data. In Proc. of ACL, pp 164–171 (1993)
6. Dekang Lin: Automatic Retrieval and Clustering of Similar Words. In Proc. of COLING/ACL, pp. 768–774 (1998)
7. Nobuyuki Shimizu, Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama, Hiroshi Nakagawa: Metric Learning for Synonym Acquisition. In Proc. of COLING, pp. 793–800 (2008)