Could we automatically reproduce semantic relations of an Information Retrieval thesaurus?

Alexander Panchenko
Center for Natural Language Processing (CENTAL), Catholic University of Louvain
Collège Erasme, 1 place Blaise Pascal, B-1348, Louvain-la-Neuve, Belgium
alexander.panchenko@student.uclouvain.be

ABSTRACT

A well constructed thesaurus is recognized as a valuable source of semantic information for various applications, especially for Information Retrieval. The main hindrances to using thesaurus-oriented approaches are the high complexity and cost of manual thesauri creation. This paper addresses the problem of automatic thesaurus construction, namely we study the quality of automatically extracted semantic relations as compared with the semantic relations of a manually crafted thesaurus. The vector-space model based on syntactic contexts was used to generate relations between the concepts of the thesaurus. We propose a simple algorithm for representing both single word and multiword terms in the distributional space of syntactic contexts. Furthermore, we propose a method for evaluation quality of the extracted relations based on the fuzzy versions of the manually constructed thesaurus. Our experiments show significant difference between the automatically and manually constructed relations: while many of the automatically generated relations are relevant, just a small part of them could be found in the original thesaurus.

Keywords
Thesaurus, Semantic relations, Vector-space model, Distributional analysis, Multiword expressions.

1. INTRODUCTION

An information retrieval thesaurus describes a certain knowledge domain by listing all its main concepts and semantic relations between them. In their simplest form thesauri consist of a list of important terms and semantic relations between them (see Figure 1). Thesauri have been used in documentation management projects for years. They were even used by libraries and documentation centers long before the computer era. This long tradition and the more recent success of the thesaurus based information systems has led to adoption of thesaurus-based techniques by the industry and to the development of international standards1.

According to Foskett [1], the main purposes to use a thesaurus are (1) to provide a standard vocabulary for indexing and searching, (2) to assist users with locating terms for proper query formulation, and (3) to provide classified hierarchies that allow the broadening and narrowing of the current request according to the needs of the user.

EuroVOC [2] is one example of a big contemporary information retrieval thesaurus: it is used for indexing documents of the European Parliament, the Office for Official Publications of the European Communities, and many other European institutions. Another well-known thesaurus is AgroVOC [3] – a multilingual, structured and controlled

---

vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains. This resource was created by the Food and Agriculture Organization of the United Nations (FAO) and has many applications all over the world.

Apart from the applications in Information Retrieval [4], the semantic information contained in thesauri and ontologies was used in solving technical problems such as Text Categorization [5], Term Extraction [6], developing Question Answering systems [7] and some others.

The traditional way of thesaurus construction involves great amount of manual labor and proved to be very time consuming and costly. Furthermore, it does not allow for an easy way to keep semantic resources updated. All these factors limit applications of thesaurus-oriented approaches. One of the solutions to this problem is to automatize thesaurus construction, as it was proposed for instance in our previous work [8]. Basically, the automatized process comprises two main steps: selecting key terms for a given domain and establishing semantic relations such as synonymy, hyponymy, and association between them. Important question concerns the quality of an automatically generated thesaurus. In this paper we investigate how similar are the automatically generated semantic relations and the semantic relations established by an expert. In our experiments we use vocabulary of a manually constructed thesaurus and try to reconstruct semantic relations between its terms by means of distributional analysis.

The paper is organized as follows. The section 2 lists some related works. We present our dataset in the 3rd section. The section 4 gives description of our method for mining semantic relations from corpus and from §4.1 to §4.4 we give details about each of its steps. Then, in section 5, we present our approach for evaluation set of automatically constructed relations and its results for our dataset. We show that while many of the automatically extracted relations make sense, the model did not recall many of the manually crafted relations. Finally we sum up the main points of this paper in section 6.

2. RELATED WORK

There has been proposed number of approaches for automatic discovering of semantic relation between words: with help of lexical and dependency patterns [9], based on Latent Semantic Analysis [10], from evidence contained in electronic dictionaries [11] or encyclopedias [12], and even from the Web link structure [13].

Yet another well-known method for discovering semantic relations between terms relies on the Distributional Hypothesis of Harris [14] which states that “words that occur in the same contexts tend to have similar meanings”. Schutze [15] proposed to represent word as a vector in a multidimensional space of all possible contexts. The spatial proximity between terms in this model indicates how similar their meanings are. There have been proposed different variations of this thesaurus construction method (e.g. [16], [17], [18] or [19]), especially in combination with clustering techniques such as in the work of Sharon [19] or Pantel and Lin [20]. We use the vector-space model based on syntactic contexts as in the work of Grefenstette [21], and extend it to deal also with multiword expressions and not only with nouns as in the original work.

3. DATASET

The dataset we are working with comprises two parts: a 20 million word corpus of political texts in French and a manually constructed thesaurus. The corpus comprise 11,386 text documents coming from a governmental institution, such as deputies’ requests to ministers, protocols of parliamentary sessions, international conventions, activity reports, texts of propositions of new laws and so on.
The thesaurus was constructed manually based on the analysis of the described above corpus. The semantic resource aims to provide vocabulary for indexing documents of a governmental institution such as a parliament, thus it comprises different terms coming from various domains (12 in our case) which are often discussed in such an institution e.g. legislation, economics, finances, international relations etc. The thesaurus contains \( n = 2514 \) concepts \( C = \{c_1, \ldots, c_n\} \) where every concept \( c_i \) is represented with \( j \) terms \( \{d_{i1}, \ldots, d_{ij}\} \) which are synonyms or quasi-synonyms. For example, the concept “Aircraft” is composed of eight terms:

\[
c_i = \{d_{i1}, \ldots, d_{ij}\} = \{ \text{Aircraft, Airship, Plane, Aerostat, Helicopter, ..., Dirigeable} \}.
\]

The terms are the key part of the thesaurus – its vocabulary, they reflect main concepts of a certain domain. The vocabulary of the thesaurus \( D \) comprises \( m = 4771 \) terms:

\[
D = \bigcup_{c_i \in C} c_i = \{d_1, \ldots, d_m\}.
\]

Most of the terms in the vocabulary (65%) are noun phrases, such as “ultra-lightweight aircraft” or “hot-air balloon”, and the rest 35% of terms are nouns, like “airplane” or “aerostat”. The concepts are organized in the hierarchy with set of 2456 hyponymy relations \( R^{NT} \). Every semantic relation \( R^{NT}_{ij} \) defines a semantic link between concepts \( c_i \) and \( c_j \) and represented with the ordered pair \( \langle c_i, c_j \rangle \). The concepts of the thesaurus are also interconnected with the set of 1530 associative relations \( R^{RT} \). Thus, the thesaurus is an oriented graph (network) \( T = (C, R) \) having the concepts of the thesaurus \( C \) as nodes, and the semantic relations between concepts \( R = R^{NT} \cup R^{RT} \) as edges.

4. CONSTRUCTING SEMANTIC RELATIONS BETWEEN CONCEPTS

Given a corpus and a set of concepts or terms, the goal of our method is to construct semantic relations between them. We use the distributional analysis [21] to construct set of semantic relations between terms of the original thesaurus. In this model every input concept is modeled as a point in the distributional space of all possible syntactic contexts. The procedure of calculating relations between the concepts involves preprocessing, indexing terms, constructing distributional space of terms, and calculation of relations between terms. The following paragraphs describe the respective steps of the proposed method.

4.1 Preprocessing vocabulary and corpus

The goal of the first step is to perform cleansing of the dataset: we use regular expressions to normalize whitespaces, remove corrupted character sequences, and some meta-information, such as document identifiers, from the texts. Also at this step we deaccent documents and terms by substituting the characters with French diacritic symbols such as “à” or “é” with their non accented equivalents.

4.2 Indexing terms

The goal of this step is to find all occurrences of the terms \( d \in D \) in the corpus and save this information in an index. In order to deal with linguistic variation and some typos we search terms with help of regular expressions. We use the Algorithm 1 to generate a regular expression for each term of the thesaurus. The procedure rely on the stemming function \( \text{Stem()} \) and the

---

\[ \text{Here and in further examples we provide the corresponding translation from French for convenience of the reader.} \]

\[ \text{We use a simplified version of the Porter stemming algorithm, which strips endings like «s», «es», and «aux» for long words.} \]
function `GetType()`\(^4\) which returns the type of an input word. The Algorithm 1 replaces every word of an input term with a regular expression pattern. Thus, the procedure replaces every article or preposition with their conjunction (line 4-5), it constructs regular expression for a regular word based on its stem form (lines 6-9). Finally, the function constructs a pattern for abbreviations by adding a special spacer after every letter of an acronym (see lines 10-17). The described procedure will transform the term “conventions internationales” (international conventions) the following regular expression:
\[
\text{\textbackslash bconvention\textbackslash w(0,3)\textbackslash s+internationale\textbackslash w(0,3)\textbackslash b}
\]
This regular expression captures both singular form “convention internationale” and plural form “conventions internationals” of the phrase. Similarly, the automatically generated regular expression for the term “modification de la legislation” (modification of legislation) will capture different pertinent variations of this term such as “modifications de la legislation”, “modification a la legislation”, or “modifications dans la legislation”.

We run the Algorithm 1 for every term \(d\) of the thesaurus and save information about every occurrence in the index record \(\langle d, doc, p_{\text{beg}}, p_{\text{end}} \rangle\), where \(p_{\text{beg}}\) and \(p_{\text{end}}\) are positions of the beginning and the end of the term in the document \(doc\). Set of all index records compose the index.

The Figure 2 shows that the terms’ frequency distribution approximately follows the Zipf’s Law \([22]\). Although, one can see that the real distribution doesn’t ideally fit the Zipf’s distribution in the area of very high- and low- frequency terms. It is mostly due to the fact that our vocabulary is just a subset of the real vocabulary of the corpus.

4.3 Constructing distributional space of terms
To construct the distributional space associated to the corpus we use syntactic dependencies between words of sentences where at least one term \(d \in D\) was found. In our experiments we used XIP natural language parser \([23]\) to produce set of syntactic dependencies \(SR\) from the corpus. Every dependency \(\langle w_1, p_{1 \text{beg}}, t, w_2, p_{2 \text{beg}} \rangle\) contains information about the syntactic relation of type \(t\) between the word \(w_1\), starting at the position \(p_{1 \text{beg}}\) and the word \(w_2\), starting at the position \(p_{2 \text{beg}}\). Some syntactic relations such as dependency between a nominal head and a determiner (e.g. \(\langle\text{the}, 0, \text{DET}, \text{helicopter}, 5\rangle\)) brings little information about the semantics of the head word. We choose 9 syntactic relations listed in Table 1 to construct the distributional space of terms. The table also indicates what syntactic relations were used in experiments of some

\[^4\text{The function use stop-lists and regular expressions. The type "articles or preposition" was defined with the 28 function words: de, du, la, le, les, des, d’, l, d, l, a, aux, et, au, en, pour, dans, par, car, dont, done, comme, que, plus, encore, entre, vers, via.}\]
other researchers. This comparison is not exhaustive, but still we can see that the most popular, and presumably giving us the best clues about meaning of a word, are the OBJ, SUBJ, and ADJMOD relations.

Table 1. Syntactic relations used to construct distributional space by A)the author, B) Piersman et al. [24], C) Hindle [25], D) Hirshman et al. [26], E) Hatzivassiloglou et al. [27], F) Lonneke [28], G) Takenobu et al. [29], F) Grefenstette [21]

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description of syntactic relation</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJMOD</td>
<td>Attaches the modifier of adjective to the adjective itself.</td>
<td>X</td>
<td>X</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>F</td>
</tr>
<tr>
<td>CONNECT</td>
<td>Links the verb of a finite clause to the grammatical word that introduces the clause.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>COORD</td>
<td>Coordination. This binary relation links coordinated elements.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DOBJ</td>
<td>This dependency attaches a deep object to the verb.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DSUBJ</td>
<td>This dependency attaches a deep subject to the verb.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>NMOD</td>
<td>Attaches a modifier to the noun it modifies</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>OBJ</td>
<td>Attaches a direct object to its verb.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SUBJ</td>
<td>Attaches the surface subject to the verb, including infinitive verbs.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>VMOD</td>
<td>Attaches a modifier of a verb to the verb itself.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DET</td>
<td>Links a nominal head and a determiner.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>APP</td>
<td>Apposition. Links two adjacent units that have identical referents</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PREPOBJ</td>
<td>Attaches a preposition to the noun or the verb it precedes.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

At this stage we have to define a distributional space and represent the terms of thesaurus in this space. The dimensions of the distributional space must be such that they let us encode the meaning of the terms, and then distinguish terms with different meanings. In our case the dimensions of the $n$-dimensional distributional space are associated with the syntactic contexts $B = \{\beta_1, \ldots, \beta_n\}$. Every syntactic context is a tuple $\langle t, w \rangle$ composed of the lemmatized word $w$ and the type of syntactic relation $t$. We derive set of syntactic contexts (features) from the set of extracted syntactic dependencies $SR$. Basically, one tuple $\langle w_1, p_1^{beg}, t, w_2, p_2^{beg} \rangle$ gives two syntactic contexts $\langle t, w_1 \rangle$ and $\langle t, w_2 \rangle$. Every term $d_i$ is represented with a vector $f_i$ in the distributional space. The feature matrix $F = (f_1, \ldots, f_m)^T$ has $m$ rows and $n$ columns, the $i$-th row of this matrix corresponds to the term $d_i$ and $j$-th column corresponds to the syntactic feature $\beta_j$.

We use the Algorithm 2 to calculate the dimensions of the distributional space $B$ and the feature matrix $F$. The majority of the previous studies describe algorithms which represent a single word or a chunk in the distributional space (e.g. [21], [24], or [29]). The main difference of our algorithm is what it can calculate representation in the distributional space of an arbitrary multiword expression (or a text segment) and not only of a single word. Basically, it calculates the distributional representation of a term as a sum of syntactic contexts of all its non-stopword, excluding the dependencies with the stopwords or linking words inside the term (see Figure 3).

The algorithm takes as input set syntactic dependencies $SR$, index $I$ containing positions of all occurrences of terms in the corpus, and the stoplists. At the first step the algorithm creates void set of syntactic contexts $B$ and void multiset $C$. An element of the multiset $C$ is a tuple $\langle d, \beta \rangle$ which maps a term $d$ and a syntactic context $\beta$. Then the algorithm incrementally fills

---

5 We adopted these descriptions mostly from the documentation of the XIP parser [23].
these two sets by checking every extracted syntactic tuple (lines 2-16). In particular, if the word \( w_1 \) from the dependency \( \langle w_1, p_{\text{beg}}^1, t, w_2, p_{\text{beg}}^2 \rangle \) belongs to the term \( d \) then we add the syntactic context \( \langle t, w_2 \rangle \) to the term \( d \). Similarly, if the term index \( I \) contains a record indicating that the word \( w_2 \) belong to the term \( d \) we add new syntactic context \( \langle t, w_1 \rangle \) to the \( d \). Furthermore, the algorithm will not add the syntactic context \( \langle t, w_{\text{context}} \rangle \) to the term \( d \) if the context word \( w_{\text{context}} \) is a part of term \( d \), or if it is a stopword (lines 12-13).

The second part of the algorithm (lines 17-21) constructs the feature matrix \( F \) from the multiset \( C \). Firstly, we set every element \( f_{ij} \) of this matrix equal to the number of times term \( d_i \) occurred with the context \( \beta_j \) (lines 18-19). Then, we normalize the feature matrix as follows (line 20):

\[
f'_{ij} = \frac{f_{ij}}{|d_i| |\beta_j|}.
\]

In the previous formula \( |d_i| \) is the number of times the term \( d_i \) occurred in the corpus and \( |\beta_j| \) is the number of times the syntactic contexts \( \beta_j \) occurred in the corpus. After the normalization every element of the feature matrix became normalized between zero and one: \( f_{ij} \in [0;1] \).

The procedure \texttt{GroupContexts()} reduces sparsity of the distributional space by merging the similar syntactic contexts such as \( \langle \text{NMOD}, 37 \text{ millions} \rangle \) and \( \langle \text{NMOD}, 71 \text{ millions} \rangle \). The procedure groups features representing dates, sums of money, ordinal numbers, real numbers and percents. Finally, in the procedure \texttt{RemoveContexts()} deletes the syntactic contexts which occurred less than \( T \) times \( \beta' \) in the corpus: \( B' = \{ \beta \in B : \beta \geq \beta' \} \). We present results of experiments with different values of this parameter in the section 5.2.

**Algorithm 2**: Calculation of feature space \( B \) and feature matrix \( F \)

**Input**: Syntactic dependencies \( SR = \{ SR_1, \ldots, SR_K \} \) for \( K \) documents, terms \( D \), term index \( I \), stop parts-of-speech \( SP \), stopwords \( SW \), allowed types of syntactic dependencies \( T \), syntactic context threshold \( \beta' \)

**Output**: \( B - \) distributional space, \( F - \) feature matrix

```plaintext
C \leftarrow \emptyset, B \leftarrow \emptyset, w_{\text{context}} \leftarrow \emptyset, w_{\text{term}} \leftarrow \emptyset;
// Calculating set of syntactic features B and multiset C
2 foreach document k in corpus do
3 foreach \( \langle w_i, p_{\text{beg}}^i, t, w_j, p_{\text{beg}}^j \rangle \in SR_k \) do
4 if \( \exists \langle k, d_i, p_{\text{beg}}^i, t, \rangle \in I : \beta_j \in [p_{\text{beg}}^i, p_{\text{end}}^i] \) then
5 \( w_{\text{context}} \leftarrow w_i \);
6 \( w_{\text{term}} \leftarrow d_i \);
7 else if \( \exists \langle k, d_i, p_{\text{beg}}^i, t, \rangle \in I : \beta_j \in [p_{\text{beg}}^i, p_{\text{end}}^i] \) then
8 \( w_{\text{context}} \leftarrow w_j \);
9 \( w_{\text{term}} \leftarrow d_i \);
10 else continue;
11 if \( w_{\text{context}} \notin w_{\text{term}} \text{ and } w_{\text{term}} \notin w_{\text{context}} \text{ and } t \in T \text{ and GetPOS}(w_{\text{context}}) \notin SP \text{ and GetPOS}(w_{\text{context}}) \notin SW \) then
12 \( \beta_i \leftarrow \langle t, w_{\text{context}} \rangle ;
13 B \leftarrow B \cup \beta_i ;
14 C \leftarrow C \cup \langle w_{\text{term}}, \beta_i \rangle ;
15 // Calculating feature matrix F
16 F \leftarrow \emptyset, \|B|,\|C|;
17 foreach \( \langle d_i, \beta_j \rangle \in C \) do
18 \( f_{ij} \leftarrow f_{ij} + 1 \);
19 Normalize(F);
20 GroupContexts(F, B);
21 RemoveContexts(F, B, \beta');
22 return B, F
```
4.4 Calculations of relations between terms
We calculate measures of semantic similarity between terms \( d_i \) and \( d_j \) with cosine between their respective vectors

\[
sim(d_i, d_j) = s_{ij} = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|}.
\]

We define set of related terms for the term \( d \) as the set of its nearest neighbors. We calculate set of relations between terms by thresholding the similarity matrix \( S \) with the threshold \( s^T \):

\[
\hat{R} = \{(t_i, t_j) : s_{ij} \geq s^T\}.
\]

5. EVALUATION
5.1 Assessment protocol
Our evaluation is based on the idea that among all possible automatically constructed thesauri \( \{(C, \hat{R}_1), (C, \hat{R}_2), \ldots\} \) the best one is the one which is the most similar to the manually constructed thesaurus \( T = (C, R) \). We evaluate quality of the automatically constructed relations with the exact and the fuzzy precision measures. The exact precision measure is defined as number of automatically extracted relations which are found in the manually constructed thesaurus, divided by the total number of extracted relations:

\[
\text{precision}^E = \frac{|\hat{R} \cap R|}{|\hat{R}|}.
\]

The original thesaurus is a hand crafted linguistic resource containing 3986 different semantic relations between the concepts. It was created by a concrete group of experts, and if another group of experts would be asked to build the same thesaurus they would created a different semantic resource. Therefore the thesaurus contains not exhaustive list of semantic links between the concepts, and the exact precision measure could tend to underestimate the real precision rate. Let us illustrate this issue on the following example: in one of our experiments the algorithm discovered that the term “foreign public act” is related to the three following terms “private international law”, “civil procedure”, “arbitration”. Meanwhile, the original thesaurus contains two different terms related to the “foreign public act”: “legal act” and “foreign legislation”. There is no overlap between the two lists and the precision of the result will equal zero.

We propose the fuzzy precision measure which addresses this problem by taking into account short paths between terms in the thesaurus. Indeed, we found that the thesaurus contains the following short transit paths between the term “foreign public act” and the automatically discovered terms:

- foreign public act → foreign legislation → branch of law → private international law
- foreign public act → legal act → course of law → civil procedure
- foreign public act → legal act → course of law → civil procedure → arbitration

To calculate the fuzzy precision score we generate set of fuzzy semantic relations \( R^{FK} \) and use it as a golden standard for evaluating quality of the automatically constructed relations. Generating set of fuzzy relations comprises the three following steps:

1. Constructing adjacency matrix \( W \) of the thesaurus \( T \) defined as follows:
2. Calculating matrix of shortest paths $P$ between concepts of the thesaurus $T$ with the Floyd’s algorithm [30]. An element of this matrix $p_{ij}$ contains length of the shortest path between the concepts $c_i$ and $c_j$.

3. Calculating set of fuzzy relations $R^{F_k}$ between terms containing pairs of terms linked with a path in the original thesaurus with length less than $k$: $R^{F_k} = \{ (c_i, c_j) : p_{ij} \leq k \}$.

In our experiments we constructed two fuzzy versions of the original thesaurus: $R^{F_3}$ and $R^{F_4}$. The first set contains 80,641 pairs of concepts which are linked by a path in the thesaurus with length less or equal than $k = 3$. The second set contains 254,441 relations; it was constructed with the maximum path length equals $k = 4$. The fuzzy precision measure is defined as number of automatically extracted relations which were found in the corresponding version of the fuzzy thesaurus, divided by the total number of extracted relations:

$$\text{precision}^{F_k} = \frac{|\hat{R} \cap R^{F_k}|}{|\hat{R}|}, k = \{3, 4\}.$$  

5.2 Results

The Table 2 presents some relations between terms of the thesaurus which were automatically extracted from the corpus with the described method. The number in brackets is the length of the shortest path in the original thesaurus $T$ between the term from the left column and the term from the right column.

<table>
<thead>
<tr>
<th>Term</th>
<th>Manually constructed</th>
<th>Automatically constructed</th>
</tr>
</thead>
<tbody>
<tr>
<td>administration of taxes</td>
<td>administration of the state</td>
<td>administration of the cadastre and the topography (2), state socio-educational center (8), public education (4), cultural institution (8), institute of hygiene and public health (7), state vineyard station (6)</td>
</tr>
<tr>
<td>admission to studies</td>
<td>school organization, education, admission to employment</td>
<td>archives of the state (9), certificate of teacher (6), program of studies (2)</td>
</tr>
<tr>
<td>medical assistance</td>
<td>medical organization</td>
<td>emergency medical services (1), medical analysis (6), medically assisted procreation (6) hygiene (6), wine institute (9), medical organization (1) medical profession (3), vaccination (5)</td>
</tr>
<tr>
<td>european election</td>
<td>election, political life, european parliament</td>
<td>legislative election (2)</td>
</tr>
<tr>
<td>unemployed person</td>
<td>unemployment, employment, employment administration</td>
<td>unemployment compensation (2)</td>
</tr>
<tr>
<td>education grants</td>
<td>school life, education</td>
<td>youth movement (11)</td>
</tr>
<tr>
<td>european community</td>
<td>european organisation, single european act, yaounde agreement, lome convention</td>
<td>european defense community (1), european atomic energy community (1), european coal and steel community (1), belgium-luxembourg economic union (2), benelux (2)</td>
</tr>
<tr>
<td>school leaving certificate</td>
<td>diploma, promotion of students, school environment</td>
<td>foreign education certificate (2)</td>
</tr>
<tr>
<td>maternity leave</td>
<td>leave, number of hours, work</td>
<td>parental leave (3), work schedule (3)</td>
</tr>
</tbody>
</table>

$^6$ We used the following parameters to generate these relations: $s^T = 0.4, \beta^T = 75$
We conducted several experiments with different values of the minimum syntactic context frequency $\beta^T \in [0; \infty]$ and the similarity matrix threshold $s^T \in [0;1]$. The figure 4(a) shows that the automatically and manually constructed relations are completely different with respect to the exact quality measure $\text{precision}^E$: the highest value of this rate is around 7%. This rate was obtained by the model keeping all the syntactic features ($\beta^T = 0$) and with similarity threshold value equals $s^T = 0.4$.

The figures 4(b) and 4(c) shows that every second (for $k = 4$) or third ($k = 3$) automatically extracted relation is present in the original thesaurus: the highest values of the fuzzy precision measures are $\text{precision}^F = 46\%$ and $\text{precision}^F = 35\%$, respectively. These scores were achieved also with the similarity matrix threshold $s^T = 0.4$, but on the distributional space composed of the syntactic contexts occurred more than 75 times in corpus: $\beta^T = 75$.

![Figure 4.](image)

6. CONCLUSION AND FUTURE WORK
Firstly, we proposed a simple method for extracting semantic relations between multiword terms, based on the distributional analysis. The method was used to reproduce semantic relations between terms of the manually constructed Information Retrieval thesaurus. Secondly, we proposed a technique for evaluating the quality of the automatically extracted relations based on fuzzy versions of the manually constructed thesaurus.

The answer to the question in the title of the article is as follows: the proposed method cannot exactly reproduce relations from the original thesaurus, but it is capable of finding pairs of terms linked with a short path in the original thesaurus. The experiments show significant difference between the automatically and manually constructed relations. Nevertheless, our observations suggest that the proposed method can discover new relevant relations between the terms. We conclude that the method could be useful in the process of automatic thesaurus construction, but its results might require moderation of an expert.

The future work will be focused on overcoming the main limitations of the method: low precision rate, need to tune the threshold parameters, and the fact that the method does not return type of the extracted relations.
7. ACKNOWLEDGMENTS
The author wishes to thank Wallonie-Bruxelles International organisation for supporting this research and professors Cédrick Fairon, Yuri Philippovich and Marco Saerens for their discerning comments and notable suggestions about this work. The author also thanks Xerox Research Centre Europe for providing their parser for French.

8. REFERENCES


