Method for Extracting Translation Correspondences from a Parallel Corpus[i]

Yu. Morozova
Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russia

Abstract – *The research paper deals with actual problems of semantic studies using methods of corpus linguistics. It overviews a new research area - distributional semantics. The method for extracting translation correspondences from a parallel corpus using context vector spaces is described. The model was tested on a parallel corpus of patent texts in Russian and French. An approximate evaluation of precision is 78%. False positive results can be explained by two main reasons. In the first place, productive syntactic transformations result in translation correspondences containing two words which are semantically equivalent but belong to different parts of speech. In the second place, words are often part of multi-word expressions and in such cases a correct single-word translation correspondence cannot be found. We propose to enhance existing models by moving from single lexemes to multi-word expressions.*

Keywords: distributional semantics, vector spaces, multi-word expressions, collocations, parallel corpora.

1 Distributional Semantics Models Overview

Distributional semantics is a field of linguistic research that aims at calculation of semantic proximity between different linguistic units using their distributional properties in large linguistic corpora. Distributional models are used in numerous research projects dealing with semantics of natural language and have a diverse range of potential and working applications. Main application areas of distributional semantics models are: lexical ambiguity resolution, information retrieval, document clustering, automatic extraction of lexicographic information (dictionaries of semantic relations, multilingual dictionaries, semantic maps of different subject areas), modeling of synonymy, document topic detection, sentiment analysis, bioinformatics.

Theoretical foundations of distributional semantics go back to the distributional methodology proposed by Z. Harris [1, 2]. Similar ideas were expressed by the founders of structural linguistics F. de Saussure and L. Vitgenstein. The theoretical basis of distributional models is the distributional hypothesis stating that linguistic units with similar distributions have similar meanings [3, 4].

Linear algebra is used as the computational instrument and as the means of model representation. First the information on linguistic units distribution is represented in the form of multidimensional vectors. These vectors constitute a matrix, in which vectors correspond to linguistic units (words or word combinations) and dimensions correspond to contexts of different sizes (documents, paragraphs, sentences, word combinations, words). When a matrix is populated from texts, semantic proximity between linguistic units can be calculated as the distance between vectors.

To compute the distance between vectors one can use various formulas: Minkowski distance, Manhattan distance, Euclidean distance, Chebyshev distance, scalar product, cosine measure. The most widely used formula is the cosine measure:

$$\frac{x \bullet y}{|x| \bullet |y|} = \frac{\sum_{i=1}^{n} x_i \bullet y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \bullet \sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (1)$$

where x_i and y_i represent frequency counts for different dimensions of the vectors.

There are many different types of distributional semantics models which differ according to the following parameters:

- type of the context (its size, left or right, ranking);
- measure used to calculate the frequency of a word in a given context (absolute frequency, entropy, mutual information etc.);
- method used to compute the distance between vectors (cosine measure, scalar product, Minkowski distance etc.);
- method used to reduce matrix dimensionality (Random Projection, Singular Value Decomposition etc.).

The most popular distributional semantics models are Latent Semantic Analysis which was designed to solve the synonymy problem in information retrieval [5], and the model of Hyperspace Analogue to Language created as the model of human semantic memory[6].

The idea of semantic vector spaces was first realized in the information retrieval system SMART[7]. Documents from a text collection are represented as vectors in a vector space. A user inquiry is viewed as a pseudodocument and

is also represented as a vector in the same vector space. The system finds n vectors of documents which are closest to the vector of the inquiry. The results are sorted by distance between vectors which reflects semantic proximity and shown to the user.

Later on the idea of semantic vector spaces was applied successfully to other semantic tasks. For example, in the research [8] a vector space was used to evaluate semantic proximity of words. The system reached the accuracy level of 92.5% on TOEFL tests to choose a synonym out of a set of words, average human result for this test being 64.5%.

At the present time there are many research projects aimed at unifying the model of vector spaces and working out a common approach to different tasks of detection of semantic relations from text corpora [9].

2 Alignment Models Overview

The task of finding single-word translation correspondences from parallel texts is part of the alignment problem and is discussed in numerous papers. In [10] an alignment is defined as an arbitrary relation between source and target language words (including correspondences of one word to an "empty" word and of one word to several adjacent words). However, the development of alignment models that are able to deal with this general model is difficult. Most often there are additional constraints to the general definition of alignment. Usually each source word is assigned to exactly one target word (including empty words). Some papers propose to add linguistic knowledge to alignment models that is used to filter out incorrect alignments.

The alignment models were first implemented by Brown and colleagues [11]. They use aligned parallel corpora and establish a translational relation between terms that occur with similar distributions in corresponding text segments.

There are two types of alignment models: heuristic models and statistical alignment models. Heuristic models are more widely used by researchers as they are easier to understand, implement and tune.

Heuristic models use a function of the similarity between words of two languages. Most often variations of the Dice coefficient are used as the similarity function:

$$dice(i, j) = \frac{2 \bullet C(e_i, f_j)}{C(e_i) \bullet C(f_j)} \qquad (2)$$

where *f, e* are the source word and the target word, *C(e,f)* is the co-occurrence frequency of *e* and *f* in the parallel corpus, *C(f)* is the frequency of *f* in the source sentences, *C(e)* is the frequency of *e* in the target sentences.

For each sentence pair a matrix with association scores between every source word and every target word is built. Then the word with the largest association score is chosen as the translation correspondence for a given word.

Within the statistical approach to alignment, the translation probability $\Pr(f_1^J \mid e_1^I)$ which describes the relationship between a source language string f_1^J and a target language string e_1^I is modeled. Statistical models depend on a set of unknown parameters Θ that is learned from the corpus. For each sentence pair (f_s, e_s), the alignment variable is denoted by $a = a_1^J$. The unknown parameters Θ are determined by maximizing the likelihood on the parallel training corpus:

$$\Theta = \arg\max \prod_{s=1}^{S} \sum_{a} p_\Theta(f_s, a \mid e_s) \qquad (3)$$

The expectation maximization algorithm is typically used to perform this maximization.

The paper [10] provides a comparison of error rate percentages for different alignment models. Heuristic models give the best result of 21.5% error rate and statistical models give the best result of 16.4% error rate for the training corpus of 0.5 K.

The papers [12, 13] introduce the cognitive approach to alignment using semantics of language units. A new grammar formalism called Cognitive Transfer Grammar (CTG) is described. The basis of CTG is composed of the proto-typical language structures, their most probable positions in a sentence, statistical data about the distributive characteristics of structures, the schemes of the complete parse of sentences.

In CTG the functional values of language structures are determined by the categorial values of head vertices. Probability characteristics are introduced into the rules of derivation in the form of the weights assigned to the parse trees.

The cognitive approach to alignment is based on the principle « from the meaning to the form ». It establishes correspondences between structures belonging to different language levels, for example : word → word, phrase structure → phrase structure, word → phrase structure, morpheme → word, morpheme → phrase structure.

The elementary structure of CTG is a *transfeme.* A transfeme is a unit of cognitive transfer, i.e. a semantic element embodied in a translatable semantically relevant language segment taken in the unity of its categorial and functional characteristics, that establishes semantic correspondence between language structures, which belong to different language levels. The types of transfemes are determined by the rank of transfemes:

- rank 1: lexemes as structural signs, i.e., a word, considered as a categorical - functional unit without taking into account the specific lexical value of this word;

- rank 2: a word combination, i.e., the syntactic structure, which consists of two and more syntactically connected words, but never a complete sentence (clause);
- rank 3: a clausal unit, i.e., dependent (subordinate) clause;
- rank 4: a sentence (either a simple sentence or the main clause of a complex sentence);
- rank 5: a scattered structure, i.e., a word group, which is characterized by a syntactic and semantic unity, but is discontinuous, i.e., between the members of the group there appear other language objects, which are not the members of this group;
- rank 0: the morphological units, which are not independent words, but which form a part of a lexeme of a source language, and in the language of transfer can be expressed by a clause and the units of other ranks.

Transfemes are represented as the rewriting rules in which a nonterminal symbol is in the left side and right sides contain the aligned pairs of chains of terminal and nonterminal symbols belonging to the source and target languages:

$$T \rightarrow < \rho , \alpha, \square > \quad (4)$$

where *T* is a nonterminal symbol (transfeme), ρ and α are the chains of the terminal and nonterminal symbols which belong to Russian and English, □ is the symbol of correspondence between the nonterminal symbols occuring in ρ and with the nonterminal symbols occurring in α. During the alignment of parallel texts on the basis of CTG the process of derivation begins from the pair of the connected initial symbols Sρ and Sα, further at each step the connected nonterminal symbols in pairs are copied with the use of two components of uniform rule.

The linguistic knowledge base described in [13] comprises the following components :

- the initial basic collection of grammar rules represented in the formalized form (CTG);
- the mechanisms of expansion and refinement of the system of rules, implemented by means of the methods of machine learning on parallel texts.

The CTG allows to automatically extract syntactical translation rules from parallel texts. Texts need to be aligned on the level of sentences and on the level of words before the rule extraction module can start working.

Different types of phrase structures are described as functional meanings together with their categorial embodiments. The transferability of phrase structures is possible when language units belonging to the same functional transfer fields (FTF) are chosen in the source and the target languages, notwithstanding the difference of their syntactic categories. The most important FTF are the following:

- Primary Predication FTF (complexes of finite verbal forms and tensed verbal phrase structures);
- Secondary Predication FTF (non-finite verbal forms and constructions, subordinate clauses comprising the finite verbal forms);
- Nomination and Relativity FTF (language structures performing the nominative functions, including the sentential units);
- Modality and Mood FTF (modal verbs and word combinations expressing modality, subjunctivity and conditionality);
- Connectivity FTF (lexical – syntactic means employed for concatenation of similar syntactic groups and subordination of syntactic structures);
- Attributiveness FTF (adjectives and adjectival phrases, nominal modifiers of different kinds);
- Metrics and Parameters FTF (language means for presenting entities in terms of parameters and values, measures, numerical information).

3 Extracting Translation Correspondences from Parallel Texts

Our research is aimed at implementing the model of semantic vector spaces to extract single-word translation correspondences from parallel texts.

The paper [14] describes a method for applying distributional semantics models to extract translation correspondences of single terms from aligned parallel texts. In general, systems extracting translation correspondences use the co-occurrence frequency of terms in the source and the target language in aligned segments as the basis for their work. The authors of [14] propose to use sentences rather than words as a minimal unit for analysis as "the primary meaning bearing unit is the utterance, the coherent expression of something meaningful by a speaker or a writer". Lexical units occurring in the same sentence are linked by syntagmatic relations and the sentence in the source language as a whole is related to the sentence in the target language by the relation of translation correspondence. Thus each word in the source sentence is related to each word in the target sentence.

In the model proposed in [14] the identification numbers of aligned regions are used as dimensions of vectors. Context vectors describing words of source and target languages are put in the same vector space. To compute the correlation between words the cosine measure is used. Words of different languages whose vectors are closest to each other are considered to be translations of each other. This approach is especially efficient when one needs to find not only the best translation but several ways to translate a term.

Within the framework of our research a test corpus of parallel texts in French and Russian aligned at the level of sentences was created. It comprises texts of scientific

patents from different areas. The volume of the corpus is 100000 word tokens. The texts were uploaded into the online corpus management system Sketch Engine [15] which provided morphological annotation of texts (lemmas, parts of speech and grammatical characteristics).

We developed a vector space model for extracting single-word translation correspondences and tried its work on the test corpus. The model is characterized by the following parameters:

- type of linguistic units: single terms;
- type of context: aligned regions;
- frequency measure: absolute frequency;
- method used to compute the distance between vectors: cosine measure.

The computer program realizing this model was implemented by Charnine M. M. [16].

Before populating the vector space we preprocessed the texts in the following way:

- words were replaces by lemmas
- the most frequent words were removed (mainly functional words such as prepositions and conjunctions)
- punctuation marks were removed.

In the result of application of the vector space model we received a list of single-word translation correspondences, for example:
moyen ("means") → средство ("means")
caractériser ("to characterize") → отличать ("to characterize")
moins ("less") → мера ("measure")
notamment ("in particular") → частность ("detail").

In many cases a translation correspondence contains two words which are semantically equivalent but belong to different parts of speech, for example:
connaître ("to know") → известный ("well-known"). Words of different parts of speech can translate each other in certain contexts but such word pairs in general should not be included in a translation dictionary. Thus the results produced by the vector model need to be filtered so that only words with the same category are left.

At the same time translation correspondences containing words of different syntactic categories provide an interesting by-product as examples of this kind correspond to productive syntactic transformations occurring during translation.

The most frequent transformations of syntactic categories for patent texts are the following:

- verbal infinitive (French) → noun (Russian)
  For example,
  traiter ("to process") → обработка ("processing").
- noun (French) → adjective (Russian)
  For example,
  crochet ("hook") → крюкообразный ("hook-shaped").
- verbal infinitive (French) → adjective (Russian)
  For example,
  connaître ("to know") → известный ("well-known").

Information about shifts of syntactic categories is a useful resource for development of a syntactic transfer module of a Machine Translation System.

On the other hand, words are often part of multi-word expressions for which the number of words is different in two languages, for example:
au moins ("at least") → по меньшей мере ("at least").
In such cases it is not possible to find a correct translation correspondence using the existing vector space model.

We envisage enhancing the current vector space model by moving from single words to multi-word expressions or collocations. By collocations we mean statistically stable word combinations. Computational linguists use different statistical measures to extract collocations from texts. As it is stated in the paper [17], the Mutual Information Measure (MI) gives the best results on average. It can be computed using the following formula:

$$MI = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)} \qquad (5)$$

where n is the first word of a collocation; c is the second word of the collocation, f(n,c) is the absolute frequency of two words occurring together, f(n), f(c) are the absolute frequencies for each single word, N is the size of the corpus in tokens.

Using MI on the texts from the test corpus we compiled a dictionary of multi-word expressions in Russian and French for the subject area of scientific patents. The following step will be to adjust the vector space model so that it can find translation correspondences of collocations from parallel texts.

4 Conclusions

The paper overviews main research areas and models of a new linguistic discipline – distributional semantics. Multi-dimensional matrices of linear algebra are used as the mathematical model, which represent a suitable formalism for computer realization. Using distributional models it is possible to automatically compile different linguistic resources on the basis of large corpora: semantic dictionaries, translation dictionaries, semantic maps of subject areas.

The present research aims at implementing distributional approach for compiling a dictionary of translation correspondences on the basis of a parallel corpus. The task of finding single-word translation correspondences from parallel texts is part of the alignment problem. The paper overviews different approaches to alignment: heuristic models, statistical alignment models, cognitive approach. Heuristic models are used most widely

as they are easy to implement. The best practical results are obtained using statistical alignment models. The cognitive approach to alignment aims at establishing correspondences between structures belonging to different language levels, for example : word → word, phrase structure → phrase structure, word → phrase structure, morpheme → word, morpheme → phrase structure. The Cognitive Transfer Grammar described in [12, 13] makes it possible to automatically extract syntactical translation rules from parallel texts.

In the result of the present research a vector space model for extracting single-word translation correspondences from a parallel corpus was implemented and tested on a parallel corpus of patent texts in Russian and French. The results of testing the model can be divided in the following groups:

- correct correspondences (semantically equivalent words of the same part of speech);
- semantically equivalent words of different parts of speech;
- fragments of multi-word expressions;
- erroneous correspondences.

Correct correspondences constitute 78% of results. Translation correspondences containing words of different syntactic categories correspond to productive syntactic transformations occurring during translation. This information can be used for development of a syntactic transfer module of a Machine Translation System. Fragments of multi-word expressions cannot be processed correctly using the current vector space model. Multi-word expressions (also called collocations) occur frequently in texts so it is necessary to include collocations into the model.

The first step consists in extracting collocations from texts in Russian and in French independently. This can be realized using statistical measures of association. The Mutual Information Measure (MI) gives the best results on average [17]. Using MI on the texts from the test corpus a dictionary of multi-word expressions in Russian and French was compiled. The next step will consist in including multi-word expressions in the vector space and finding translation correspondences between them by computing the distance between vectors.

5 References

[1] Harris Z. S. “Papers in Structural and Transformational Linguistics”. D. Reidel., 1970.

[2] Harris Z. S. “Mathematical Structures of Language”. Interscience Publishers John Wiley & Sons, 1968.

[3] Sahlgren M. “The Distributional Hypothesis. From context to meaning”; Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Vol. 20, Issue 1, 33-53, 2008.

[4] Turney P. D., Pantel P. “From frequency to meaning: Vector space models of semantics”; Journal of Artificial Intelligence Research, Vol. 37, 141-188, 2010.

[5] Landauer Th. K., McNamara D. S., Dennis S., Kintsch W. “Handbook of Latent Semantic Analysis”. Lawrence Erlbaum Associates, 2007.

[6] Lund K., Burgess C. “Producing high-dimensional semantic spaces from lexical co-occurrence”; Behavior Research Methods, Instruments & Computers, Vol. 28, Issue 2, 203-208, 1996.

[7] Salton G. M. “The SMART Retrieval System: Experiments in Automatic Document Processing”. Prentice-Hall, 1971.

[8] Rapp R. “Word sense discovery based on sense descriptor dissimilarity”; Proceedings of the 9th MT Summit, 315–322, 2003.

[9] Turney P. “A uniform approach to analogies, synonyms, antonyms and associations”; Proceedings of COLING, 905–912, 2008.

[10] Franz Josef Och, Hermann Ney. “A Systematic Comparison of Various Statistical Alignment Models”; Computational Linguistics, Vol. 29, Issue 1, 19-51, 2003.

[11] Brown P. S., Cocke V., Della Pietra F., Della Pietra F., Jelinek R., Mercer, Roossin P. “A statistical approach to language translation”; Proceedings of the 12th Annual Conference on Computational Linguistics, 1988.

[12] Kozerenko E.B. “Parallel Texts Alignment Stratagies: The Semantic Aspects”; Informatics and its Applications, Vol. 7, Issue 1, 82-89, 2013.

[13] Kozerenko E.B. “Syntactic Transformations Modelling for Hybrid Machine Translation”; Proceedings of WORLDCOMP’11, 875-881, 2011.

[14] Sahlgren M., Karlgren J. “Automatic Bilingual Lexicon Acquisition Using Random Indexing of Parallel Corpora”; Journal of Natural Language Engineering (Special Issue on Parallel Texts), Vol. 11, Issue 3, 2005.

[15] Web site for Sketch Engine: http://www.sketchengine.co.uk/

[16] Kuznetsov, I.P. Elena B. Kozerenko, Mikhail M. Charnine. “Technological peculiarity of knowledge extraction for logical-analytical systems”; Proceedings of ICAI’12, WORLDCOMP’12, July 18-21, 2012, Las Vegas, Nevada, USA. CRSEA Press, USA, 2012.

[17] Zaharov V.P., Hohlova M.V. “The Analysis of Efficiency of Statistical Methods of Collocations Detection for Texts in Russian”; Computational Linguistics and Intelligent Technologies (International Conference “Dialog”), 2010.

[i] The work is supported by the Russian Foundation of Basic Research, grant 11-06-00476-a.