VILNIUS TECH
Vilnius Gediminas Technical University

Mokslas – Lietuvos ateitis / Science – Future of Lithuania
ISSN 2029-2341 / eISSN 2029-2252
2021 Volume 13, Article ID: mla.2021.15034, 1–8
https://doi.org/10.3846/mla.2021.15034

Information technologies and multimedia
Informacinės technologijos ir multimedija

SEMI-AUTOMATIC ONTOLOGICAL ALIGNMENT OF DIGITIZED BOOKS PARALLEL CORPORA

Algirdas LAUKAITIS [1*], Neda LAUKAITYTĖ[2]

[1]*Vilnius Gediminas Technical University, Vilnius, Lithuania*
[2]*Vilnius University, Vilnius, Lithuania*

Received 30 May 2021; accepted 23 June 2021

Abstract. In this paper, we present a method for general ontology management integration with an alignment of digitized books paraphrase corpus, which have been compiled from bilingual parallel corpus. We show that our method can improve ontology development and consistency checking when we add semantic parsing and machine translation to the process of general knowledge management. Additionally, we argue that the focus on one's favorite books gives a factor of gamification for knowledge management process. A new formalism of semantic parsing ontological alignments is introduced and its use for ontology development and consistency checking is discussed. It is shown that existing general ontologies requires much more axioms than it is currently available in order to explain unaligned content of books. Proactive learning approach is suggested as part of the solution to improve development of ontology predicates and axioms. WordNet, FrameNet and SUMO ontologies are used as a starting knowledge base of paraphrase corpus semantic alignment method.

Keywords: ontological alignment of corpora, alignment of digitized books, machine translation, natural language processing.

Introduction

An alignment of parallel bilingual corpus is an important task for machine translation systems, and it is now widely used for the extraction of translation patterns. Yet, when it comes to an alignment of corpus consisting of translated fiction books, the existing alignment algorithms give low precision results (Laukaitis et al., 2011).

As an example of the limitations of existing alignment models, we can consider some random sentence from Stanislaw Lem's novel Solaris (this was the first novel from which we started our project) and its Russian, English, Lithuanian translations (see Figure 1). If we look at Polish-Russian alignment pairs, then it is easy to see that statistical machine translation (SMT) can align almost each word in the sentence to its counterpart in translation. The state of the art SMT models like the alignment template translation model (Och & Ney, 2003) or hierarchical phrase-based translation model (Chiang, 2007) can align such sentences with high precision. But, when we run SMT system on Polish-English pair of sentences, then, we get a set of alignments where only few words are aligned. Yet, both English and Polish sentences semantically express the same meaning.

As this example suggest the often used source–channel approach (Brown et al., 1993) or log-linear models (Och & Ney, 2004) must be complemented with general knowledge about the surrounding world in order to explain different interpretations of the same meaning in these books. In this paper, we propose a solution to this alignment problem that complements existing SMT models. We suggest creating additional alignments between words (or phrases) in a sentence of the book and predicates (or axioms) in general ontology. In order to illustrate this idea, consider the English word *"head"* from our example (Figure 1). It can be connected to the Polish word *"twarz"* (in English *"face"*) by using SUMO (Niles & Pease, 2001) axiom defined in knowledge interchange format (KIF) language:

```
(=>
    (instance ?FACE Face)
    (exists (?HEAD)
        (and
            (instance ?HEAD Head)
            (part ?FACE ?HEAD))))
```

*Corresponding author. E-mail: *algirdas.laukaitis@vilniustech.lt*

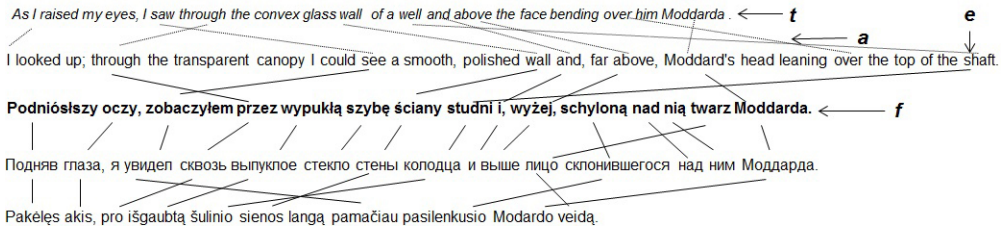

Figure 1. Example of lexical alignment. Values of variables (*e, f, t, a)* are inputs in our stochastic ontology alignment model (examples of full text alignments can be found in *https://github.com/algirdaslaukaitis/BooksAlignment*)

This axiom can be read as: if an object 1 is an instance of *"face"*, then there exists another object 2 such that the other object 2 is an instance of *"head"* and the object 1 is a part of the other object 2.

This axiom can explain why different words in sentence can express the same state of the world. Then, it looks like a good idea to use ontologies in order to improve alignments that we get from more figurative translation. But there are several obstacles to implementing this approach:

1. Ambiguities in dictionary (i.e. WordNet, FrameNet) between ontology concept and natural language word;
2. Sparsity of the ontology axioms and predicates;
3. Consistency between all concepts, axioms and predicates in the ontology;
4. Almost all semantic parsing research has been done only for English language.

As an illustration of the ambiguity problem we can consider the same example between English word *"head"* and the Polish word *"twarz"*. If we look at the WordNet dictionary for the word 'head' we will find that there are 33 synsets that are mapped manually to 25 SUMO concepts. The choice of which one to use is a considerable challenge.

The problem of sparsity of an ontology axioms can be realized from the fact that there are 2550 axioms in SUMO ontology (SUMO claims to be one of the most axiomatized upper ontologies).

Our experiments revealed that we are able to explain less than 1% of unaligned open class words in our set of translated sentences using these axioms. This means that the set of axioms in existing ontologies is underdeveloped for any practical use when we talk about machine understanding of fiction books.

Ontology consistency problem in our research context we define as inability to check logical conflicts for new entries in the ontology. For example you can enter new concept, predicate or axiom in SUMO ontology but there is no inference engine that will check if new entry is inconsistent with existing ontology records. Similarly, WordNet and FrameNet ontologies fully rely on judgment about concepts and predicates consistency on human editor.

As to the fourth consideration, we don't know of any system that can semantically parse two languages with the same upper ontology. The only way to overcome this hurdle in our research was to translate foreign language books back into English and then parse these translations with English language parsers. Fortunately, state-of-the-art automatic translation systems, especially for Russian language, can translate without almost any loss of semantics. This allows us to transform the task of bilingual language semantic alignment into paraphrase corpus semantic alignment.

Thus, the solution to these problems and the main thesis of this paper is that we need:

1. A method that integrates ontology development with such activity like book reading or translation;
2. Ontology development must be done using inference engine and a set of verifiable goals;
3. We need natural language semantic parsing to be independent from which national language we use;
4. Proactive learning must be part of the whole learning framework.

The rest of this paper is organized as follows: Section 2 describes the general process and algorithms of the multilingual corpus semantic alignments. Suggested corpus semantic alignment algorithm can be seen as an extension of the commonly accepted alignment algorithms in statistical machine translation. Our contribution is that we suggest the use of new set of hidden variables that reflects mapping between translation lexicon and concepts in general otology. Additionally, we argue that in order to have semantic parsing for any non-English language, we can use current state-of-the-art translation system to translate sentences to English and then use English language semantic parser.

In Section 3, we describe the various components of a paraphrasing pattern. Here, we extend formal definition of synchronous context-free grammar in order to integrate semantic graph that links words in a paraphrase. As its name suggest, the purpose of paraphrase pattern is to generate a set of paraphrases that have the same logical structure.

In Section 4 we describe the various elements of user interaction model. Our view is that existing ontologies are of poor quality and too small in order to find semantic alignments. Then, we need an interaction model that implements several requirements for ontology manager engagement in order to develop ontologies by reading books in parallel.

In Section 5 we present evaluation of the suggested method. We evaluate alignment quality on 863 paraphrased books. Additionally, we analyze the effect of automatically induced patterns on translation quality.

1. Related work

There are several areas of research that are relevant to the present paper. We will review each in turn.

First, there is statistical machine translation (SMT) that provided foundation for our machine learning algorithms. The early works in this area (Berger et al., 1996; Brown et al., 1993) suggested the use of expectation maximization approach in order to get Viterbi alignments between words. Later works focused on an important improvement of the use of phrases instead of just single words as the main elements of the statistical translation model (Och & Ney, 2003; Marcu & Wong, 2002). These models can robustly perform alignments on the bilingual corpus which usually represents a technical (literal) translation. But when we use a corpus of fiction books, these algorithms frequently give sparse and erroneous alignments (Laukaitis & Vasilecas, 2008; Laukaitis et al., 2011).

Natural language semantic parsing is the second area of research that influenced our work. Early attempts in semantic parsing tried to induce semantic parsers from small annotated corpus. Zelle and Mooney (1996) suggested to use inductive logic programming to learn database queries from natural language sentences. 250 samples where used in their research on US geography data base. It was been shown that in a narrow domain with up to 20 predicates we can achieve 75% precision when queering databases with 5 tables. In (Kwiatkowski et al., 2011) probabilistic CCG grammars where used to induce semantic parser from 880 samples. 88.6% precision has been reported on the geography data base Geo880. These two papers among many others show weaknesses and strengths in an attempt to induce semantic parser form annotated sentences. High precision parsers and high quality induced knowledge base is definitely the strength of these methods. On the other hand it is difficult to create sample base for learning and there is requirement to keep narrow domain in order to learn useful parser. Our contribution to this semantic parsing approach is that we suggest to create semantically annotated corpus by reading fiction books. In our research we found that semantic alignment can bring some gamification to the process of sentence annotation and that in turn can increase volume of manually annotated data.

There are few, if any, systems that try and are capable semantically parse arbitrary text from books. This is because there still remain major challenges labeling natural language words and phrases with logical concepts and predicates and to combine these concepts and predicates into a coherent logical form. Several recent approaches aimed to lift limitation of manually annotated corpus and tried to learn a semantic parser without annotated logical forms. Berant et al. (2013) used 596M assertions that link predicates and 41M entities from Freebase knowledge base in order to infer logical predicates from natural language utterance.

Even on a bigger scale (Mitchell et al., 2018) tried to infer predicates and logical expression from the whole Internet. In our research we found that these approaches are not well suited for semantic alignment of fiction books because huge amount of automatically induces erroneous predicates and axioms that can lead to big misalignment errors. Nevertheless we use ideas form these approaches to suggest some logical expressions for ontology manager, who then can decide whether or not to enter logical expressions into the general ontology.

The main stage in semantic parsing is a correct mapping between words/phrases and logic entities and it's called word sense disambiguation (WSD) problem. Usually WSD systems that are based on supervised learning (Navigli, 2009) requires large amounts of hand-labeled data. An alternative to the supervised learning approach can be dictionary as a graph approach. In this case we can use some graph analysis algorithm (like personalized PageRank (Agirre et al., 2014)) to rank word concepts by their graph properties. In this paper we take a hybrid approach. At first we use personalized PageRank approach. We increase probabilities of graph vertices if group of words can be connected using axioms and predicates from ontology. Then, we ask book reader to confirm or reject disambiguation results, and then we use this information in process of supervising learning.

Automatic induction of axioms and predicates can generate many irrelevant or erroneous ontology elements. That is why we need confirmation from ontology manager to select relevant ones. The research in the area of active learning and crowdsourcing can help efficiently build up queries for ontology manager to ease selection process of relevant ontologies entities. There have been few research papers that investigated active learning use NLP tasks, such as text classification (McCallum & Nigam, 1998; Tong & Koller, 2001) or information extraction (Thompson et al., 1999; Settles & Craven, 2008). We found that it is straight forward to adopt theses active learning strategies for book alignment and word sense disambiguation. It is less clear how to use active learning in case when we want to query for ontology axioms and predicates. In this paper we suggest to add reader utility function that leverage machine query with user preferences.

2. Ontology based alignments

Informally, we define ontology based alignments as an explanation in some formal language for differences in English language paraphrases. As an example of our approach we can analyze, word *"glass"*, (Polish *"szybę"*), phrase *"transparent canopy"* , and *"window"* (Lithuanian *"langą"*) from our sentence example in Figure 1. We would like to have a system that, like a human, is able to explain why

seemingly different words like "glass", "transparent canopy" and "window" refer to the same object in the scene.

One of the many possible logical explanations for these references can be seen in Figure 2. There, all words are interconnected by using SUMO ontology concepts and axioms. With the axiom:

```
(=>
    (instance ?W Window)
    (attribute ?W Transparent))
```

we can explain why word *"transparent"* relates to the word *"window"*. On the other hand there is no axioms in standard SUMO ontology that explains why we can relate the same object in scene by the words *"glass"* and *"window"*. In this case we would like for the alignment system to suggest a possible axiom or allow ontology designer to enter it manually. Below, we can see one of the few axioms for this alignment.

```
(=>
    (instance ?Glass Glass)
    (exists (?Window)
    (and
        (instance ?Window Window)
        (part ?Glass ?Window))))
```

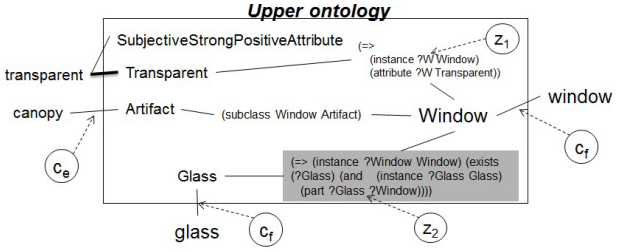

Figure 2. Ontology alignment that let us to connect words *"transparent"*, *"canopy"*, *"window"* and *"glass"* into a single semantic graph

We can see, from this informal introduction that, in addition to existing variables in traditional SMT model, we need several sets of variables in order to create probabilistic model for ontology alignments. In Figure 3 we present a graphical model for these random variables.

Our analysis we start from the set of variables e, f, a_{ef}. Variables e, f represent pair of sentences where f is translation of English sentence e. A hidden alignment variable a_{ef} describes a mapping between words in this pair of sentences (see Figure 1). The relationship between these variables in the SMT alignment model (Brown et al., 1993) is given by

$$P(f \mid e) = \sum_{a_{ef}} P(f, a_{ef} \mid e). \quad (1)$$

In order to find variables a_{ef} we solve optimisation problem:

$$\widehat{a}_{ef} = \arg\max_{a_{ef}} P(f, a_{ef} \mid e), \quad (2)$$

where: $\widehat{a}_{ef}$ is an alignment that has the highest probability and it is called the Viterbi alignment. If we look at Figure 1 then the words in Polish sentence can be interpreted as

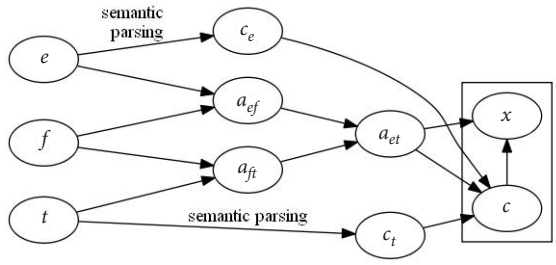

Figure 3. Graphical model for random variables

variable f, words in English sentence can be interpreted as variable e and straight line links can be interpreted as variable a_{ef}.

One of the ideas that we would like to suggest in this paper is to do semantic analysis for non-English language sentences by translating them to English and then aligning translation with original sentence. Thus, variable t represents translation to English language from f using automatic translation system. We interpret it as a paraphrase sentence to original English sentence e. Variable a_{et} describes a mapping between words in f and t. We can use the same expression (2) to find lexical alignments between f and t. Variable a_{et} describes a mapping between words in e and t. In order to find $\widehat{a}_{et}$ we can use information about a_{ef} and a_{ft} using following expression:

$$\widehat{a}_{et} = \arg\max_{a_{et}} P(t, a_{et} \mid e, \widehat{a}_{ef}, \widehat{a}_{ft}). \quad (3)$$

There are several general semantic parsers for English language that we were able to use in our project. Each of these parsers generates set of labels over English phrases. We use variables c_e and c_t to mark these labels and we used the following parsers in order to find values for these variables:

1. *SEMAFOR* system that annotates sentence phrases with FrameNet concepts;
2. *StanfordNLP* named entities and sentiments recognition system;
3. *NLTK WordNet* synsets disambiguation system;
4. *NLTK SUMO* ontology concepts disambiguation system.

3. User interface patterns

Process of book semantic alignments can be a motivating factor for an ontology manager in order to test and improve ontology. But, for ontology manager to stay engaged in semantic alignment process, we need user-friendly interface, and in order to build one, we need a data structure that integrates natural language processing tools in our possession. In this section we suggest the data structure that, in its essence, is an English phrase from original book and it's English paraphrase that we get after automatic translation into English with at least one word replaced with variable defined by ontology class. Let look at phrase *"I looked up"* from our example in Figure 1 and its paraphrase *"I raised my eyes"*. We can replace word

"looked" with FrameNet concept *"Scrutiny"*, word *"up"* with concept name *"Direction"*, *"raised"* – *"Motion"* and *"eyes"* – *"Body_parts"*. Then the pattern that we get after replacement operation will be: *"I <Scrutiny> <Direction>" <=> "I <Motion> my <Body_parts>"*.

This is an erroneous pattern if we analyze it as a logic statement (i.e. Scrutiny and Direction is not equivalent to Motion and Body_parts). Logical error can come from errors in semantic parsers and from incompleteness and inconsistencies in ontologies. It is up to ontology manager to find it and make corrections, and our research goal is to suggest the most relevant corrections.

Ontology labels are only one part of the user interface pattern. The full pattern description has the following three parts (see Figure 4):

1. Syntax information: 1) synchronous context-free grammar (i.e. aligned parsing trees): a set of rewrite rules with aligned pairs of right-hand sides, 2) POS labels, 3) dependency grammar;
2. Semantic parsing data consists from WordNet, FrameNet and Sumo ontology labels, predicates and rules. Labels are partitioned into three subsets: labels that are consistent with Viterbi alignments, labels that contradicts Viterbi alignments and labels that are partly consistent with Viterbi alignments;
3. A set of generated paraphrases. Some paraphrases can have confirmation or rejection labels from book reader.

In order to define first two parts of semantic alignment pattern, we introduce an extended synchronous context-free grammar (CFG). In an ordinary synchronous CFG (Chiang, 2007) the basic elements are text transformation rules with aligned pairs of right-hand sides.

$$X \rightarrow < \gamma, \alpha, \sim >, \quad (4)$$

where: γ and α are strings of terminals and nonterminals, X is a nonterminal and $\sim$ is a one-to-one correspondence between nonterminal occurrences in γ and nonterminal occurrences in α.

The implementation of our extension of synchronous CFG is straightforward and it is defined as:

$$X \rightarrow < \gamma, \alpha, \sim, \beta >, \quad (5)$$

where: extension β means a set of connections to ontology concepts in semantic subgraph.

Informally, the first part of the pattern is an alignment between branches of two parsing trees, where the first tree is a parsing tree from original English sentence, and the second one is paraphrased tree that we receive by taking translation of foreign language sentence to English language using automatic translation system. The root of this parsing tree represents nonterminal from parsing grammar. The labels of the tree leaves are either the words from sentence or the words replaces by ontology concepts. All other nodes of parsing tree are labeled by POS labels or by the concept names from ontology. The second part of the pattern represents a set of connected ontology concepts that have been used to label parsing tree nodes. These connections are expressed either by predicates or by axioms from ontology. In order to be included into the pattern these elements must be verified by ontology manager.

4. Proactive learning

Active learning deals with finding optimal query that maximizes informativeness of correct answer from oracle. The simplest example from ontology alignment framework would be to ask book reader to choose words in a pair of sentences. Ontology alignment informativeness of these words can be based on our ability to apply new knowledge to other sentences in the corpus.

In this section we suggest to integrate ontology alignment and pattern induction algorithms into proactive learning framework. Proactive learning is a generalization of active learning with purpose to relax assumptions that oracle (in our case, book reader) is foolproof, always answers or insensitive to costs. Additionally to those assumptions we consider new metric of query interestingness to oracle. The main idea is to generate set of sentences that reflects the same semantics on given abstraction level. The user can manipulate abstraction level for each concept.

In order to implement this idea we focus on several proactive learning scenarios that are designed to explore different oracle preferences. We express these scenarios from book reader perspective and from active learning algorithm perspective. These scenarios are shown in Figure 5 as UML Use Case Diagram.

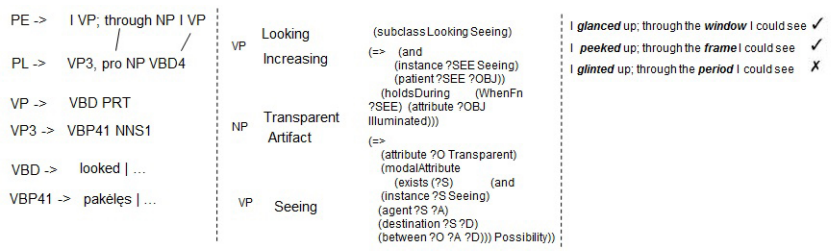

Figure 4. Example of semantic alignment pattern

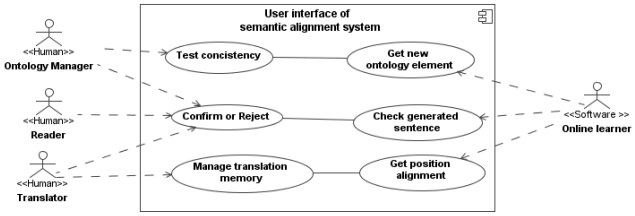

Figure 5. Alignment processes during which the system can query for labels

In the first scenario, we focus on the single oracle where the main objective of the oracle is to improve his upper ontology by expressing book meaning in ontology formal language. We name this scenario as consistency checking. We say that ontology is consistent with text in a book if it is able to explain alignments that are not explained by word/phrase based statistical alignments.

The second scenario is intended for readers who are not familiar with knowledge representation in some formal logic language. In this scenario, we focus on the third part of semantic pattern. Semantic alignment system generates series of new sentences using method that we presented in the section above. The book reader just accepts them or rejects them. These acceptation/rejection labels then are used by online learner to modify probability distribution of semantic pattern items.

The left part of the figure shows activities that human user can use with semantic alignment system. On the right side of the figure we show related activities that correspond to queries posed by proactive learner. We can see from these activities that we need not just to choose the most informative sample for the learning process but, also we need to find the way how to present these samples to user. So rather than looking for instances to sample, as in standard active and proactive learning, we focus on dialog planning between oracle and machine where objective is to maximize information gain *V(S)* and oracle expected utility function *U()*:

$$D_i = \max(k_i \cdot E(V(S)) + h_i \cdot U()), \quad (6)$$

where: D_i is the decision score for reader *i*. More details on this model can be found in (Settles & Craven, 2008).

5. Evaluation

We begin our evaluation of this framework for semantic alignment by defining Alignment Error Rate (AER) (Och & Ney, 2003). AER can be defined on the sentence-to-sentence and word-to-word levels and it requires a manually aligned set of "sure" (used for measuring recall) and "possible" (used for measuring precision) links (referred to as *S* and *P*).

$$AER(A,P,S) = 1 - \frac{|P \cap A| + |S \cap A|}{|A| + |S|}. \quad (7)$$

We evaluate our semantic alignment method using two other methods (one from our previous study) against which we compare quality of corpus alignment. The first *"hunalign"* is the method suggested by (Varga et al., 2007). We chose this method because we think resources that we have for Lithuanian language are similar to resources in their study. As a second method for estimating alignment quality we have used a method *"bookalign"* from our previous study on bilingual alignment of books corpus (Laukaitis et al., 2018). Table 1 shows the statistics of the corpus used to evaluate these methods.

There were two questions that we considered, namely, whether the suggested ontology management framework is useful for improving alignment precision of paraphrased books and how alignment quality depends on a number of ontology queries that user must answer.

In order to answer these questions, we conducted the following experiment. We created the bilingual English-Lithuanian corpus of 859 books. Then we used all three methods to align this corpus. The error rates that we received after this step are shown in the Table 2. We can see that from first column of Table 2 that error rates are significant when we don't use proactive learning.

Using proactive learning framework the system generated queries after the first alignment iteration. Once the query has been answered, new alignment iteration started. The next columns in Table 2 show error rates obtained after these steps. We iterated this alignment loop until we answered 35 queries. It is clear from examining the results that all three methods improved performance after each an-

Table 1. Corpus statistics for alignment quality assessment (see *https://github.com/algirdaslaukaitis/BooksAlignment*)

	Sentences	Words	Vocabulary
English	6978447	658214931	137564
Lithuanian	6815472	624467239	418711

Table 2. Fiction books sentence-to-sentence alignment error rates for different methods

Anchor p. No.	0	2	10	15	20	35
hunalign	0.54	0.52	0.42	0.24	0.23	0.19
bookalign	0.47	0.45	0.37	0.21	0.11	0.07
semalign	0.43	0.41	0.32	0.15	0.09	0.06

swered query. What is particularly interesting, however, is that number of queries required to align books in order to get error rate below 0.1 can differ significantly for each book. Nevertheless, the number of 35 queries appeared as a limit, after which all books can be aligned with acceptable quality.

Conclusions

In this paper we have suggested a model for semantic bilingual corpus alignment. Our method allows us to extend statistical alignment model by adding a new semantic layer of concepts. It has been shown that in this model for semantic parsing task it is possible to transform bilingual corpus into English language paraphrase corpus using machine translation systems.

Then we have described the semantic template induction process that uses this paraphrase corpus. The main novelty here is that we focus on unaligned words and phrases in paraphrase corpus. These unaligned words and the fact that we are dealing with one's favorite books give us a puzzle game on how to link these words using semantic templates. Formal logic knowledge is required in order to play this game. But, we believe that it is possible to improve this game by designing dialog between human and machine. In this game the reader would explain why unaligned words in paraphrases express the same meaning and computer automatically induce required template.

We would like to note about our choice of the ontology. In this paper we experimented with SUMO ontology and its map to WordNet. Many other ontologies are available and we thing that paraphrasing corpus can be one of the tools to test them and integrate into a knowledge base. We believe that it is possible to manage general ontology and fiction books corpus by a convenient selection of the proactive learning method. Moreover, this new learning method could be used to generate new set of axioms for general ontology using, for example, unaligned concepts from fiction books.

References

Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, *40*(1), 57–84. https://doi.org/10.1162/COLI_a_00164

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1533–1544). Association for Computational Linguistics.

Berger, A. L., Della Pietra, V. J., & Della Pietra S. A. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*(1), 39–72.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., & Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, *32*(2), 201–228. https://doi.org/10.1162/coli.2007.33.2.201

Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., & Steedman, M. (2011). Lexical generalization in CCG grammar induction for semantic parsing. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *24*(18), 1512–1523.

Laukaitis, A., & Vasilecas, O. (2008). Multi-alignment templates induction. *Informatica*, *19*(4), 535–554. https://doi.org/10.15388/Informatica.2008.229

Laukaitis, A., Plikynas, D., & Ostasius, E. (2018). Sentence level alignment of digitized books parallel corpora. *Informatica*, *29*(4), 693–710. https://doi.org/10.15388/Informatica.2018.188

Laukaitis, A., Vasilecas, O., Laukaitis, R., & Plikynas, D. (2011). Semi-automatic bilingual corpus creation with zero entropy alignments. *Informatica*, *22*(2), 223–224. https://doi.org/10.15388/Informatica.2011.323

Marcu, D., & Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *10*, 133–139. https://doi.org/10.3115/1118693.1118711

McCallum, A., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. In *Proceedings of the International Conference on Machine Learning* (pp. 359–367). Morgan Kaufmann.

Mitchell, T. M., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Mishra, B. D., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., … Welling, J. (2018). Never-ending learning. *Communications of the ACM*, *61*(5), 103–115. https://doi.org/10.1145/3191513

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 10. https://doi.org/10.1145/1459352.1459355

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. *Proceedings of the International conference on Formal Ontology in Information Systems*, *2001*, 2–9. https://doi.org/10.1145/505168.505170

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–15. https://doi.org/10.1162/089120103321337421

Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, *30*(4), 417–449. https://doi.org/10.1162/0891201042544884

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1070–1079). Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613855

Thompson, C. A., Califf, M. E., & Mooney, R. J. (1999). Active learning for natural language parsing and information extraction. In *Proceedings of the 16th International Conference on Machine Learning* (pp. 406–414). Morgan Kaufmann Publishers.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, (2), 45–66.

Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., & Tron, V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies in the Theory and History of Linguistic Science*, *4*(292), 247. https://doi.org/10.1075/cilt.292.32var

Zelle, J. M., & Mooney, R. J. (1996). Learning to parse database queries using inductive logic programming. *Proceedings of the National Conference on Artificial Intelligence*, *2*, 1050–1055.

LYGIAGRETAUS SKAITMENINIŲ KNYGŲ RINKINIO DALINIS AUTOMATINIS SUGRETINIMAS, NAUDOJANT ONTOLOGIJAS

A. Laukaitis, N. Laukaitytė

Santrauka

Straipsnyje pateiktas bendrosios ontologijos valdymo metodas naudojant parafrazių rinkinius, gautus iš grožinės literatūros knygų. Straipsnyje pateiktas metodas gali pagerinti tolesnį ontologijos plėtimą ir loginio nuoseklumo patikrinimą. Šio metodo funkcionalumas grindžiamas dviem esminėmis technologijomis: semantine teksto analize ir automatiniu kompiuterio vertimu. Svarbus pateikto metodo aspektas – žaidimo elementų naudojimas valdant bendrąsias ontologijas. Šis aspektas užtikrinamas tuo, kad ontologijų valdymo procesas glaudžiai susietas su grožinės literatūros kūriniais. Straipsnyje pateiktas naujas ontologijų suderinimo formalizmas. Tyrimų rezultatai parodė, kad esamos bendrosios ontologijos turi būti papildytos kur kas didesniu kiekiu aksiomų, nei yra šiuo metu, kad būtų galima paaiškinti semantinį nesugretintų parafrazių ekvivalentiškumą. Papildomai straipsnyje pasiūlytas proaktyvus mokymosi metodas, leidžiantis pagerinti ontologijų kūrimo procesą. „WordNet“, „FrameNet“ ir SUMO ontologijos naudojamos kaip pradinės žinių bazės, siekiant pagerinti semantinio sugretinimo metodą.

Reikšminiai žodžiai: tekstų sugretinimas, ontologijų kūrimas ir naudojimas, automatinis mašininis vertimas, natūralios kalbos apdorojimo algoritmai.