October 5, 2015

RESEARCH STATEMENT
of Dr. Alisa Zhila

Broadly defined Dr. Zhila's area of research is **computational linguistics** and **natural language processing**. More specifically, her main research interest lies in **open information extraction** from text and its applications to **text quality evaluation**, in particular, text informativeness. Departing from these areas, she has also worked on the problems of **human opinion collection** for evaluation of subjective aspects of text. Another direction of Dr. Zhila's research is the mapping of the output returned by open information extraction systems onto **RDF data representation model**. In parallel to her main focus, Dr. Zhila has also worked on **semantic similarity measure** and **system log classification**.

Dr. Alisa Zhila's research contributions to the field of computational linguistics and natural language processing are described below grouped into several parts according to their topic or place where the work was conducted. This includes her research accomplishments while she was pursuing her PhD thesis (2011-2014), as well as the explorative part of her work at various employments. In all the research projects, Dr. Zhila has either played the leading research role or conducted them entirely on her own.

1. Open Information Extraction

The main focus of Dr. Zhila's research has lain on Open Information Extraction (OIE), which is an area of Computational Linguistics and Natural Language Processing aimed at detection of potentially informative fragments in arbitrary text and their extraction in semi-structured form.

In 2011 she started her research in OIE as a PhD student in the Center for Computing Research of Instituto Politecnico Nacional, which is one of the top technical universities in Mexico. Her research in Open Information Extraction took several directions.

Open Information Extraction based on Part-of-Speech Sequences

Previous research in OIE had introduced approaches based on such techniques of natural language analysis as syntactic parsing and syntactic chunking. The problem with this type of text processing is that syntactic parsing is computationally and resource costly and slow. Moreover, it is not available for a wider range of languages. Dr. Zhila introduced a novel approach to OIE based only on patterns of Part-Of-Speech (POS) sequences [9]. The first contribution of Dr. Zhila's work is that the introduced approach reduces the processing time significantly, e.g., up to 100 times on a single machine with one-thread processing as compared to full syntactic parsing. Second, it makes the OIE paradigm available to a large variety of languages because POS-taggers are available for far more languages than syntactic parsers. Third, Dr. Zhila has showed that the OIE approach based on POS patterns performs better on the arbitrary unedited texts from the Web than the approach based on syntactic chunking because the former is more robust to grammar incoherencies encountered in arbitrary text from the Web [4]. Now Dr. Zhila's method for OIE is considered one of the major approaches to OIE by the computational linguistic community.

Open Information Extraction for Spanish Language

Another direction of Dr. Zhila's research that made significant contribution to the field is OIE for Spanish language. Indeed, she was the first to introduce a method for OIE for another language rather than English. Dr. Zhila has introduced a method for transfer of the basic OIE algorithm based on POS-tags to languages with fixed word order and designed an algorithm for Spanish language [6]. She has implemented the OIE system for Spanish called ExtrHech, which is available to the community through public software repositories. As an additional contribution, Dr. Zhila has gathered several annotated datasets for experiments and system evaluation for OIE in Spanish and English [2, 9]. Before this effort, there had been no publicly available datasets for OIE evaluation in any language.

Applications of Open Information Extraction

OIE can also serve as a tool for a variety of end-user tasks. Dr. Zhila has researched its application to automatic text quality measure, in particular, such aspect of text quality as text informativeness. Automatic text quality measure is a very important research area because the quality of texts available on the Web varies significantly. The goal of this area is to ensure the delivery of only the most informative texts to a user. Dr. Zhila was the first to conduct experiments in OIE application to text informativeness measure on real internet texts and to show statistically significant positive correlation of extraction density and informativeness [8]. This work served as a basis for further research on OIE application to text quality evaluation and has been cited by a number of authors.

Named-Entity Driven Open Information Extraction

The most recent direction of Dr. Zhila's work in OIE is the introduction of a novel method for OIE driven by named entity recognition. Previously, OIE methods employed language pre-processing of the entire text and then scanned for fragments to extract through the entire text again. The idea of named-entity driven OIE proposed by Dr. Zhila is that instead of looking through all text, the method looks for potentially informative fragments only for the sentences containing a named entity. Detection of named entities in text is a much faster procedure than syntactic parsing or POS-tagging. Therefore, by applying OIE techniques only to selected sentences, the OIE method is sped up several times against the state-of-the-art OIE approach. Additionally, the named entity driven approach reduces the error rate by 32%. The trade-off of the speed-up is that this method ignores the information in the sentences that have not been selected at the pre-processing stage. The work is submitted to print [1].

OIE2RDF

Although OIE finds some direct applications as in automatic text quality and informativeness measure, most of other applications, such as knowledge base population or semantic database search, require the extraction output to be converted into a particular data representation model. One of the most wide-spread data representation model is the RDF model.

Dr. Zhila has recently been working on the conversion of the output of OIE systems into the RDF model. She has addressed such issues as the design of a property vocabulary, correspondence between the components of extracted tuples and the components of RDF statements, and, most importantly, the transformation of complex multicomponent tuples describing complex relations and phenomena, e.g., reported speech or an "inner object", into

standard RDF triples [3]. This result provided a bridge between two communities, enabling the results of natural language processing to be used in knowledge engineering.

2. Collection of Human Opinion on Subjective Text Characteristics

Collection of human opinion is a necessary process to elaborate a standard dataset against which any results of automatic text processing are evaluated in the areas that involve a subjective opinion, such as text quality, text readability, and text informativeness. Design of opinion collection methodologies adequate for computational linguistics purposes is always a hard task and often is tailored for a particular research.

Within her work on OIE application to automatic informativeness evaluation [8], Dr. Zhila designed a methodology for collecting human opinion on informativeness of a text based on simultaneous comparison of various items that was implemented by her co-author. In the further work, Dr. Zhila compared this methodology with the Likert questionnaire methodology and showed the superiority of the simultaneous item comparison method for research and evaluation in the area of automatic text quality measure. The main contribution of this work is that Dr. Zhila showed that presenting various questions for one item at the same time as done in Likert questionnaire affects subject's objectivity about each evaluated aspect [5].

3. Semantic Similarity Measure

This section describes the research that Dr. Zhila conducted while working as an intern in Microsoft Research, Redmond, WA.

Dr. Zhila invented a method for measuring semantic similarity degree between word pairs that was superior in performance than the previous state-of-the-art method. Measuring of semantic similarity is a very difficult task in computational linguistics because unlike syntax and morphology, meaning of a word is not "encoded" by its form. Yet despite its complexity, this task is crucial for text understanding and, consequently, for a variety of applications such as semantic text analysis, semantic search, and language learning applications.

The main contribution of this work is the introduction of a novel method for similarity measuring based on supervised machine learning and a combination of various semantic, lexical, and ontological features. In this work Dr. Zhila conducted an analysis of potentially useful features, implemented the corresponding components for feature engineering, determined the most appropriate learning settings given a large scale of the data, trained a classifier, and evaluated the results. The performance of Dr. Zhila's method yielded statistically significant 9% improvement in accuracy as compared to the previous best system [7]. The contribution of this work is impactful on the research community and it has been largely cited since its publication in 2013. Moreover, an extension of this method has been patented by Microsoft Research [10].

4. System Log Classification

This work was conducted when Dr. Zhila worked as an intern at Yahoo, Sunnyvale, CA. She worked in Event2Resolution group and her role was to consult the team on possible approaches to system log classification by the level of importance or emergency.

Dr. Zhila conducted statistical analysis of a large (several gigabytes) dataset of logs. She has determined that given the settings, the entropy measure that had been used to estimate how unexpected a log event was, was equivalent to the *tf-idf* measure used in computational

linguistics. She has also determined the most significant character patterns for log importance evaluation.

Jointly with her supervisor George Chen, Dr. Zhila designed and implemented a system for log annotation by human expert annotators. She collected and processed a dataset of human annotated data for the log classification task.

Further, she designed and implemented a supervised log classification system that used character pattern matches and other lexical features. Dr. Zhila conducted a series of experiments on log classification that showed the feasibility of the supervised classification approach at industrial scale.

References

[1] **Alisa Zhila,** Alexander Gelbukh, Helena Gomez-Adorno. Fast Named Entity Driven Open Information Extraction with Shallow Semantic Interpretation. *Information Sciences.* **2015** *(Submitted Sept. 22, 2015).*

[2] **Alisa Zhila**, Alexander Gelbukh. Open Information Extraction from Real Internet Texts in Spanish Using Constraints over Part-Of-Speech Sequences: Problems of the Method, Their Causes, and Ways for Improvement. *Revista Signos. Estudios de Lingüística*, 90, vol. 49. **2016** *(In print, March 2016).*

[3] **Alisa Zhila,** Elena Yagunova, Olga Makarova. Bringing The Output of Open Information Extraction to The RDF/XML Format: A Case Study. *Poster at International Conference on Knowledge Engineering and Semantic Web* (KESW-2015), **2015**.

[4] **Alisa Zhila.** Open Information Extraction based on Constraints over Part-of-Speech Sequences. PhD Thesis. **2014**.

[5] **Alisa Zhila,** Alexander Gelbukh. Informativeness and Objectivity of Texts on the Web. *Poster at Tapia Celebration of Diversity in Computing*, **2014**.

[6] **Alisa Zhila**, Alexander Gelbukh. Open Information Extraction for Spanish Language based on Syntactic Constraints. *In Proceedings* ACL (Student Research Workshop) 2014, pp. 78-85**, 2014.**

[7] **Alisa Zhila**, Scott Yih, Chris Meek, Geoffrey Zweig and Tomas Mikolov. Combining Heterogeneous Models for Measuring Relational Similarity. *In Proceedings* HLT-NAACL 2013, pp. 1000-1009, **2013**.

[8] Christopher Horn**, Alisa Zhila**, Alexander Gelbukh, Elisabeth Lex. Using Factual Density to Measure Informativeness of Web Documents. *In Proceedings* NoDaLiDa-13, pp. 227-238, **2013**.

[9] **Alisa Zhila**, Alexander Gelbukh. Comparison of Open Information Extraction for Spanish and English. *In Proceedings* Dialogue-2013, 12, vol. 1, pp. 794-802, **2013**.

[10] Wen-tau Yih, Geoffrey Zweig, Christopher Meek, **Alisa Zhila**, Tomas Mikolov. Relational similarity measurement. *US 20140249799 A1*.**2013**.