Dictionary and pattern-based recognition of organization names in Russian news texts

Valery Solovyev, Rinat Gareev, Vladimir Ivanov, Sergey Serebryakov, Natalia Vassilieva

HP Laboratories
HPL-2013-14

Keyword(s):

Named entity recognition; knowledge-based event extraction;

Abstract:

This paper describes a part of the event extraction system which has been developed in collaboration with HP Labs Russia. The domain of input texts is business news feeds. One of the most important event participant types is 'Organization'. This paper is focused on the problem of organization names recognition in Russian news texts. Two approaches have been implemented. The first is dictionary-based. We propose an algorithm to make a dictionary from a set of legal body full names gathered from a government registry. The main problems with the dictionary matching are incorrect stemming and significant fraction of ambiguous names among dictionary entries. The second recognition approach is based on usage of local context clues and internal name words. These words constitute patterns which are intrinsic to organization names. These patterns enable recognition of non-dictionary names. We propose an algorithm to derive such patterns from the original dictionary.

External Posting Date: February 21, 2013 [Fulltext] Approved for External Publication
Internal Posting Date: February 21, 2013 [Fulltext]

Procedia Information Technology & Computer Science

00 (2013) 000-000

3rd World Conference on Information Technology 2012

Dictionary and pattern-based recognition of organization names in Russian news texts

Solovyev V.D. [a] *, Gareev R.M. [b], Ivanov V.V. [c], Serebryakov S.V. [d], Vassilieva, N. S. [d]

[a]Kazan Federal University, 18 Kremlyovskaya St., Kazan 420008, Republic of Tatarstan, Russian Federation
[b]Institute of Informatics AS RT, 36a Levobulachnaya St., Kazan 420012, Republic of Tatarstan, Russian Federation
[c]National University of Science and Technology "MISIS", 4 Leninskiy pr., Moscow 119049, Russian Federation
[d]Hewlett-Packard Laboratories, 1 Artillerijskaya str, St.-Petersburg 191104, Russia Federation

Abstract

This paper describes a part of the event extraction system which has been developed in collaboration with HP Labs Russia. The domain of input texts is business news feeds. One of the most important event participant types is 'Organization'. This paper is focused on the problem of organization names recognition in Russian news texts. Two approaches have been implemented. The first is dictionary-based. We propose an algorithm to make a dictionary from a set of legal body full names gathered from a government registry. The main problems with the dictionary matching are incorrect stemming and significant fraction of ambiguous names among dictionary entries. The second recognition approach is based on usage of local context clues and internal name words. These words constitute patterns which are intrinsic to organization names. These patterns enable recognition of non-dictionary names. We propose an algorithm to derive such patterns from the original dictionary.
Keywords: Named entity recognition, knowledge-based event extraction;

Selection and/or peer review under responsibility of Prof. Dr. Dogan Ibrahim.

1. Introduction

In collaboration with HP Labs Russia we have developed an event extraction system for Russian language, in a manner similar to the original one intended for English. Both systems are intended for real-time text processing and use knowledge-based approach proposed in [8] with dictionaries and patterns. The domain of input texts is business news feeds. One of the most important event

* ADDRESS FOR CORRESPONDENCE: Karbysheva, 63/1, apt.17, Kazan, 420101, Russia
E-mail address: maki.solovyev@mail.ru / Tel.: +7- 919-691-0489

participant types is 'Organization' which includes companies as well as non-commercial institutions like government agencies. This paper is devoted to the recognition of organization names in Russian news texts.

We have assembled the dictionary of legal body names using open web catalogues. Its size is about 2.3 million entries. We have exploited a UIMA (Unstructured Information Management Architecture) as a basic framework which provides a set of common NLP components such as tokenization and sentence splitting. The dictionary matching component was built upon the ConceptMapper component [16]. Extraction rules are represented by means of TextMARKER component [10]. In this article we describe the architecture of the event extraction system, methods proposed for the dictionaries organization and optimization as well as our future work in these directions.

2. Problem statement and context of the work

Fast retrieval of precise information about current business events is especially important for financial markets, which is particularly sensitive to news [13]. For example, influence of 'Mergers & Acquisitions' events on markets is shown in [14]. Ordinary information sources in Web – news feeds – usually provide only rough categorization without capabilities to select specific event types that traders and investors are interested in.

A lot of research work is devoted to event extraction for English. Recent surveys and classification of existing approaches can be found in [6, 8]. However, only a few of them deal with business domain. A strategy of the most significant events detection is proposed in [11], however, it does not provide event types. The ACE program has marked out 33 event types. ACE does not oriented solely to business events, therefore, does not fully cover this domain. There are only 4 event types related to business. For example, there are no events like rising stock prices or company announcements.

An approach using handcrafted text patterns and rules [1, 8] is a mainstream in commercial information extraction systems. Machine learning techniques proposed in [9] are mostly of research interest so far. Thus supervised learning methods need huge manually annotated corpora, which are hard to produce. At the same time they are not proven to show better performance than systems with handcrafted rules [9]. Unsupervised learning techniques are much slower and cannot provide the user with higher quality than supervised techniques. For instance, it may take about 15 minutes to process a single query, because a system proposed in [15] uses syntax parser. Several tests of the system show that less than 50% of its answers are judged to be feasible.

In comparison with English there is much less NLP tools for Russian language, including those for information extraction. Comparing the Russian and English text processing system we discovered a set of crucial differences between them such as higher importance in part of speech parsing and shallow syntax parsing. For example, name "Газпром" (which is a common term referring to Open Joint Stock Company Gazprom) may occur in text in different forms: "Газпрома", "Газпрому" etc. which complicates dictionary matching. One of the famous NLP-toolkits for Russian, the AOT, includes POS-tagger but it is not well documented and has no technical support. The two most prominent information extraction systems for Russian are the OntosMiner [3] and RCO [4]. The key feature of both systems is that they present an environment allowing the user to manually create descriptions and patterns of desired events. Therefore users need to both know a lot about the domain and have a strong linguistic background. We propose a system that allows automatic extraction of business events using exhaustive domain-specific dictionaries and elaborated rules which do not require much user effort.

A similar approach is described in [2], but this system is aimed at extracting facts of illegal border crossings. The significant difference is that in a business domain we need to identify different legal

body names. The fact that those names have a high variety of possible expressions in news texts makes extraction much more complex. Company names recognition is a part of a system represented in [12]. But the system deals with a restricted set of companies, the "NASDAQ-100". Our system is aimed at recognizing company names of any kind for its dictionary which contains more than 2.3 million entries.

The system workflow includes 3 stages (Figure 1): document retrieval from news feeds, document pre-processing and an event extraction resulting in filling event-argument slots with corresponding text spans. A prototype of the system allows extracting 5 types of events: Company Acquisitions, Company Announcements, Person Announcements, Person-Position Assignments and Resignations. Each event type has a set of obligatory (e.g. company-acquirer) and non-obligatory slots (e.g. date, place). We use a standard text processing pipeline including following components: tokenizer, sentence splitter and word stemmer. Further we will attach POS-tagger, morphology parser (which is particularly useful for Russian) and shallow syntactic analyzer (chunker) to the pipeline. Full syntactic analysis is too slow and expensive for real-time extraction. Those components are managed by UIMA platform.

Templates for event extraction are built upon production-rules with condition-parts (constraining texts to be matched) and action-parts (representing actions to carry out when conditions from a condition part are fulfilled). All the templates are represented in TextMARKER notation. Dictionaries and gazetteers intended for automatic text annotation are also important components. They are extensively used for extraction of entities which can be used in business events – companies, persons, geo-political entities, temporal expressions, currency units, codes and amounts as well as other numerical expressions. We describe those components of the system precisely.

3. Basic dictionary-based recognition

We treat a dictionary-based recognition task as the following: given a list of strings (dictionary) find all their occurrences in an input text. Each dictionary entry may consist of several tokens: words, numbers, punctuation marks, etc. Tokens may be separated by whitespaces. Any sequence of

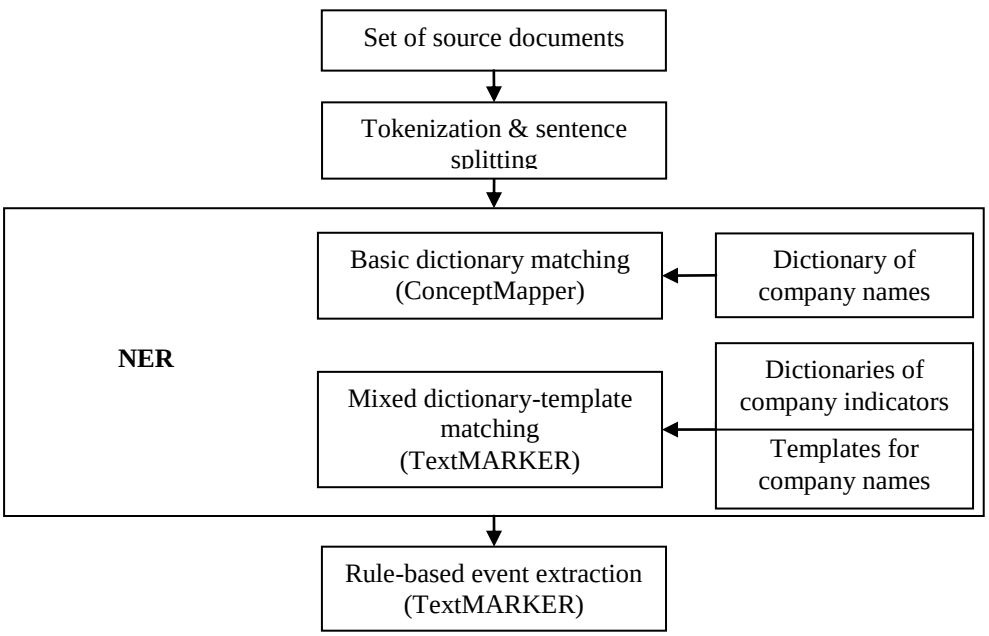

Figure 1. Event extraction pipeline

whitespace characters between two adjacent tokens in a dictionary entry or an input text is always considered as a single whitespace. Russian language is an inflectional one so stemming is applied on all Russian word tokens of dictionary entries and an input text prior to the matching process.

We have assembled the dictionary of legal body names from web catalogues. Each catalogue contains a lists of organizations registered in a particular federal district. After aggregating the results we have got 2.7 million unique names. However, full legal names are rarely used in news; hence a certain post-processing is required to enrich the original dictionary with more concise name variations. To solve this task we have implemented the algorithm which is based on the following principles:

- Derive a name variation from name substring which is surrounded by quotation marks. A nested quoted substring is also a quite frequent case, e.g. ‘ООО "Компания "Роскомплект"’.
- Refuse name variations which are equal to any of the following: common language word, person first name, geographical name.

As a dictionary of common Russian words we have used a morphological dictionary from the OpenCorpora project [7]. As a dictionary of geographical names we have used the Russian part of Geonames dump. A dictionary of people’s first names contains about 20000 entries. The result dictionary (R) contains about 2.3 million entries. The algorithm is shown in Figure 2.

4. Mixed dictionary and pattern-based recognition

The dictionary-based approach gives quite high recall and low precision. The possible reason of a large amount of false positives is incorrect stemming of some organization names which yields the same stem with a frequently used word or name of other type (e.g., person, geoname). The dictionary-based approach has the following disadvantages:

- Usage of large dictionary leads to high memory consumption.
- Dictionary has to be updated periodically.
- All names in dictionary derived from name lists are spelled in Russian. However, in Russian news feeds foreign company names are typed mostly in English. Recognition of organization names requires a dictionary of company names in English.

To overcome these drawbacks we also implemented a different approach which does not rely on

```
Input: D – the source dictionary of legal names, CW – the common words dictionary,
GN – geographical names dictionary, PN – person first names dictionary.
Output: R – the expanded names dictionary
begin
   R = ∅ - the result set
   for each d ∈ D do
         R = R ∪ {d}
         d = UnQuote(d)
         while d is not null do
               R = R ∪ {d}
               d = UnQuote(d)
   AMB = ∅ - the set of ambiguous names
   for each r ∈ R do
         if IsAmbiguous(r) then AMB = AMB ∪ {r}
   R = R \ AMB
   return R
end
function UnQuote(name)
begin
   if name contains quoted subsequence qs then
         return qs
   else return null
end
function IsAmbiguous(str)
begin
   return NOT ( (str ∈ CW) OR (str ∈ GN) OR (str ∈ PN) )
end
```

Figure 2. Dictionary preparation algorithm

lists of complete organization names. Instead it exploits the natural language indications that a certain span of text designates an organization.

Precise detecting of entity name boundaries is out of this paper scope. Proper names always start with a capitalized word. To distinguish an organization's proper name from a name of another type we can use features of the name's external context as well as internal features of the name itself. We have extracted indicators from the dictionary described in the previous section and defined 3 types of indicators:

(1) Common words and phrases usually preceding the organization name and denoting its kind. Examples: 'компания' (company), 'фонд' (foundation).

(2) Known acronyms designating a form of organization or its kind. These acronyms often represent intrinsic parts of official names. Examples: 'ЗАО' (closed joint stock company), 'ООО' (limited liability corporation).

(3) Words frequently used in multi-word organization names. Examples: 'финанс' (finance), 'инжиниринг' (engineering). In [5] these words are called 'trigger words'.

In order to derive indicators (1, 2) we have collected phrases outside the quote marks (see UnQuote function in Figure 2). Then we sorted them by frequency and retained indicators with the frequency above threshold 10. So we ended up with about 550 indicators of type (1) and 2000 indicators of type (2). To get trigger words (3) we have implemented algorithm shown in Figure 3.

After running this algorithm we have got about 2700 patterns. Given these indicators we can transform the original dictionary into several dictionaries consisting of separable indicators, unique name parts and the set of patterns which represent different ways of name composition in text. These patterns allow prediction of organization names which are not covered by the original dictionary but are expressed similarly.

5. Conclusion and future work

We described possible approaches to complex named entity recognition in a real-world knowledge-based event extraction system: a "huge-dictionary" recognition and a "mixed dictionary and pattern-based" recognition. These approaches complement each other. The second one is more flexible because it allows us to match names even if they do not exist in the dictionaries. Both approaches should be evaluated using common measures: precision and recall. We also proposed algorithms for transforming a huge dictionary into a set of smaller dictionaries and rules.

Acknowledgements

```
   Input: D – the subset of the dictionary containing only entries without quotes
   Output: P – a set of patterns consisting of trigger words
   Parameters: PatternFreqThreshold – a pattern frequency threshold (=20 in experiments)
   begin
      OWN = {d ∈ D | d is one-word name} - set of one-word names from D
      D = D \ OWN
      PM = ∅ - multiset of patterns
      for each d ∈ D do
            for each w ∈ OWN and (d contains w) do
                  p = d with replacement of w by '*'
                  PM = PM ∪ {p}
                  D = D \ {d}
      return P = {p ∈ PM | multiplicity(p) > PatternFreqThreshold}
end
```

Figure 3. Pattern generation algorithm

The authors would like to thank HP Labs Russia for funding and useful discussions with lab employees, specifically Sergey Serebryakov and Natalia Vassilieva.

References

[1] Borsje, J., Hogenboom, F., and Frasincar F. Semi-automatic financial events discovery based on lexico-semantic patterns. *In Int. J. Web Eng. Technol.*, 2010, pp 115–140.

[2] Du, M., von Etter, P., Kopotev, M, Novikov, M., Tarbeeva. N., and Yangarber, R. (2011). Building Support Tools for Russian-Language Information Extraction. *I. Habernal and V. Matousek (Eds.): TSD 2011, LNAI 6836,* 2011, pp 380–387.

[3] Efimenko, I.V., Khoroshevsky, V.F., and Klintsov, V.P. OntosMiner Family: Multilingual IE Systems. *In Proc. of International Conference SPECOM-2004.* [In Russian].

[4] Ermakov, A.E., and Pleshko, V.V. Semantic interpretation in text analysis computer systems. *Information technologies*, 2009, 155(7). [In Russian].

[5] Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., Wilks, Y. University of Sheffield: Description of the LaSIE System as Used for MUC-6. *In Proc. Message Understanding Conference*.

[6] Grishman, R. Information Extraction: Capabilities and Challenges. *Notes prepared for the 2012 International Winter School in Language and Speech Technologies Rovira i Virgili University.* Tarragona, Spain, 2012.

[7] Granovsky, D., Bocharov, V., Bichineva, S. Open corpus: principle of operation and prospects. In Proc. of XIIIth All-Russia Scientific Conference “Internet and Modern society” (IMS 2010), 2010. [In Russian].

[8] Hogenboom, F., Frasincar, F., Kaymak, U., and de Jong, F. An overview of event extraction from text. *In Proc. of Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011), 779,* 2011, pp 48–57.

[9] Indurkhya, N., and Damerau, F.J. *Handbook of Natural Language Processing (2nd ed.).* Chapman & Hall/CRC, 2010.

[10] Kluegl, P., Atzmueller, M., and Puppe, F. TextMarker: A Tool for Rule-Based Information Extraction. *Proc. Unstructured Information Management Architecture UIMA, 2nd UIMA@GSCL Workshop. Conference of the GSCL Gesellschaft für Sprachtechnologie und Computerlinguistik.*

[11] Liu, M., Liu, Y., Xiang, L., Chen, X., and Yan, Q. Extracting Key Entities and Significant Events from Online Daily News. *IDEAL 2008, LNCS 5326,* 2008, pp 201–209.

[12] Micu, A., Mast, L., Milea, V., Frasincar, F., Kaymak, U. Semantic Knowledge Management: an Ontology-based Framework. *IGI Global, Chapter Financial News Analysis a Semantic Web Approach,* 2008, pp 311–328.

[13] Mitchell, M.L., and Mulherin, J.H. The Impact of Public Information on the Stock Market. *Journal of Finance*, 1994, 49(3), pp 923–950.

[14] Rosen, R.J. Merger Momentum and Investor Sentiment: The Stock Market Reaction to Merger Announcements. *Journal of Business* 2006, 79(2), pp 987-1017.

[15] Sekine, S. On-demand information extraction. *In Proc. of Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL-06)*. Edmonton, Canada, 2006.

[16] Tanenblatt, M., Coden, A., and Sominsky, I. The ConceptMapper Approach to Named Entity Recognition. *In Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.