Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-119

Latvian Newswire Information Extraction System and Entity Knowledge Base

Pēteris PAIKENS[a,1]
[a] *University of Latvia, Institute of Mathematics and Computer Science*

Abstract. This paper describes an information extraction system designed for obtaining CV-style structured information about publicly mentioned persons, organizations and their relations by analyzing newswire archives in the Latvian language. The described text analysis pipeline consists of morphosyntactic analysis, NER and coreference resolution, and a semantic role labeling system based on FrameNet principles. We also implement an entity linking process, matching the entity mentions in each document to an entity knowledge base that is initially seeded with authoritative information on relevant people and organizations. The accuracy of automated frame extraction varies depending on specifics of each frame type, but the average accuracy currently is 53% F-score for frame target identification, and 61% for frame element role classification. The currently targeted volume of text is the total archives of Latvian newspapers, magazines and news portals, consisting of about 3.5 million articles.

Keywords. Information extraction, knowledge base, text summarization

Introduction

Newswire archives contain a huge wealth of information that has been once gathered, verified and published, but is scattered among many separate articles of unstructured natural language text. There is a large demand for extracting this knowledge in a structured and summarized manner, and this is also an important business area for a number of news broker companies. A particular niche of structured information is the profiles of important people and companies. For some languages and locations, the need is well served by open resources such as Wikipedia, but for others, including Latvia, this coverage is not sufficient and there is a market need for providing such data. This is currently done by LETA, the largest Latvian news agency, providing profiles of some 20,000 people and 2,000 organizations. The raw data of news articles is digitized and well accessible, and current technology supports effective search and retrieval of relevant documents, but creating and maintaining profile data still is very labor intensive and requires a significant time investment. This time cost limits the coverage of such profiles to a fraction of all people mentioned in news, and restricts the frequency of reviewing and updating those profiles.

As this type of information can be automatically extracted from article text by state of art information retrieval approaches, a research project was started by University of Latvia IMCS together with LETA, the largest Latvian news agency, with the goal of

[1] Corresponding Author: Pēteris Paikens, University of Latvia, Institute of Mathematics and Computer Science, Raiņa bulvāris 29, Riga, Latvia, LV-1459; E-mail: peteris@ailab.lv.

researching these methodologies, adapting them to the Latvian language and target domain, and building a prototype for a pilot project of extracting profile data about publicly mentioned persons and organizations, as well as their relations, from newswire archives in Latvian.

The described text analysis system is designed to provide news analysts with 'fact candidates' about such entities, linking to the primary sources of those facts for clarification – if this can be done with a sufficient accuracy, it allows to summarize larger amounts of data than is manageable by people using common search techniques. The structured data of relations between people and organizations can also be used for journalist analysis of indirect relations, when represented as a graph in tools that allow to visualize and explore such data. The currently targeted volume of text is the total archives of Latvian newspapers, magazines and news portals, consisting of about 3.5 million articles.

In the first section, we describe the main research problems encountered, and the relevant previous research on solving those problems. Section 2 describes the conceptual and technical architecture of the developed information extraction system. Section 3 provides details on the representation chosen to model the relevant domain facts. Section 4 describes the process of linking entity mentions discovered in text to the appropriate real world entities. Finally, we provide some conclusions and discussion of future work.

1. Problem Description and Related Work

Information extraction is a currently unsolved problem in computational linguistics, and an active area of research. While some of the required components are well-researched, many of them still require improvements to be suitable for practical usage even for well-resourced languages such as English. The main active research issues are the actual semantic data extraction phase, the abstract meaning representation model, and entity linking to the appropriate real world entities.

Implementing information extraction for the Latvian language added extra challenges in developing or adapting tools for the more general text processing stages. The morphosyntactic analysis and named entity recognition parts of this system are a separate problem that is described shortly in the next section, and with more detail in the cited publications.

1.1. Newswire Information Extraction Task

There are recently started projects for other languages with similar aims – the closest such project is NEWSREADER [1] that uses a similar methodology for aggregating notable newswire events, with their current analysis focus on the financial domain of public companies. While that research is still ongoing and was not published before our system development was well underway, their approach is very relevant and offers solutions to potential subtasks, e.g. for scaling the processing [2] if we would want to analyze the corpora in real time or move to larger corpora than only Latvian newswire.

In addition, there are multiple projects that also attempt information extraction from newswire corpora, most notably Europe Media Monitor [3], but they target a different problem scope and the main overlap with this paper is in entity identification.

1.2. Meaning Representation Model

As we need to represent the domain information in a structured manner, the choice of meaning representation determines both the scope of facts that the system will be able to describe, as well as the level of detail and nuance that it will attempt to capture from text. The classic approaches for modeling this data include relational databases and the linked data approach using Resource Description Framework models.

For the purposes of this system, we have chosen to model the domain knowledge according to FrameNet principles [4], as described in section 3. The main reason for this choice was that it is closer to the fact representation as it occurs in natural language sentences; and the advantages of other approaches can be obtained by further automated transformations of this data to RDF or specific database formats.

A notable new relevant approach has been recently published – Abstract Meaning Representation [5], which would potentially be valuable for this use case, but currently still needs more research and tool development for automated text analysis to this representation.

1.3. Semantic Role Labeling

The key component of the text analysis system is the step of mapping the identified text morphosyntactic structure and entities to the semantic representation. We treat the core part of profile data extraction as a semantic role labeling problem, annotating sentence tokens with the frame target and frame element information according to the chosen representation.

The current state of art systems for performing this step, as measured on corpus of Semeval2007 shared task, are LTH [6] and Semafor [7]. Those algorithms are of general purpose and can be adapted to a variety of languages, annotation paradigms and text domains, and were also tested in practice on Latvian data. During our research, we developed a novel, separately described method [8] based on decision tree learning that achieves a comparable accuracy, and also gives a possibility for manual rule review and improvement that is well suited for the properties of this project –preexisting domain knowledge and small number of frame types that makes manual rule review feasible.

There is also significant research on extracting such data by fixed sentence patterns and regular expressions, which we did not consider in depth as such approaches have limited coverage in languages with variable word order such as Latvian.

1.4. Entity Linking

The relevant sub problem of entity linking is the task of matching named entities identified in documents to their real-world counterparts, identifying new entities and resolving ambiguities for multiple people with the same name. A related problem also is cross-document coreference resolution and event coreference linking, which is not currently handled but is a topic for future work

Current related research on entity linking includes Wick et al [9], Han et al [10] and Stoyanov et al [11], based on which we developed an entity linking module tuned for the needs of this project as described in section 4.

2. System Architecture

The main parts of the system are its text analysis pipeline and the entity knowledge base. The text analysis pipeline consists of morphosyntactic analysis [12], [13], named entity and coreference identification [14], and a semantic role labeling system [8]. The latter two components were implemented for the Latvian language specifically for this project, and the morphosyntactic and NER layers were adapted for newswire text domain by creating additional training data and tuning statistical models.

After this document analysis, the entities found in each document are mapped to an entity-based knowledge base, as shown in Figure 1, and appending the newly identified facts. Afterwards, the facts identified in each separate document (often duplicates) are summarized for further applications.

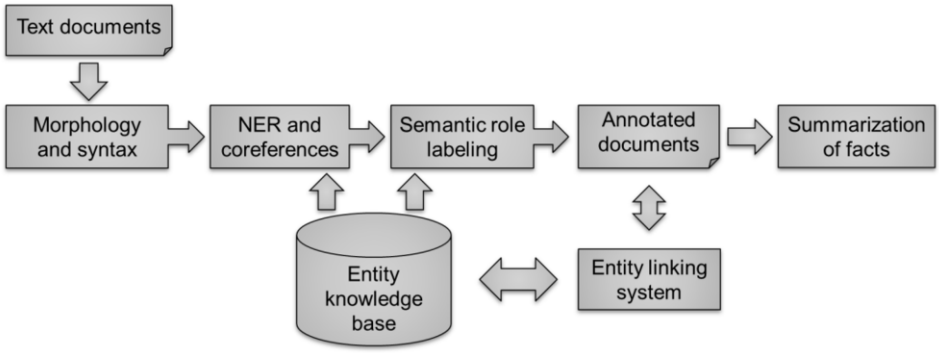

Figure 1. Analysis process flow.

The technical architecture implements each annotation layer as a separate software module capable of running independently and with multiple concurrent copies, suitable for batch processing of large corpora. Data interchange between the modules is done either in columnar tab-delimited format as used in historical CONLL and Semeval shared tasks or a custom JSON format that includes the entity details and semantic frame labeling over the original sentences.

3. Fact Representation and Entity Knowledge Base

For the purpose of analyzing biographical data, we have chosen a narrow subset of English FrameNet – 26 frames – and adapted the frame details for both the Latvian language and targeted domain.

A key challenge was the actual adaptation of semantic frame models. It was not straightforward, as the original FrameNet frames significantly vary in granularity, and domain-specific needs mandated adjustments and additions to the original frames. The currently proposed ontology, shown in Figure 2, stores the identified semantic frames as predicates linking together multiple entities, and allows summarizing/merging multiple frames with identical or overlapping information.

The implementation treats all types of entities as equal, and all information that is particular only to some entities (e.g., people) is stored as separate types of semantic frames. Thus, the entities are reduced to their identities and alternative names, the type and a set of semantic frames that describe this entity and link it to other entities.

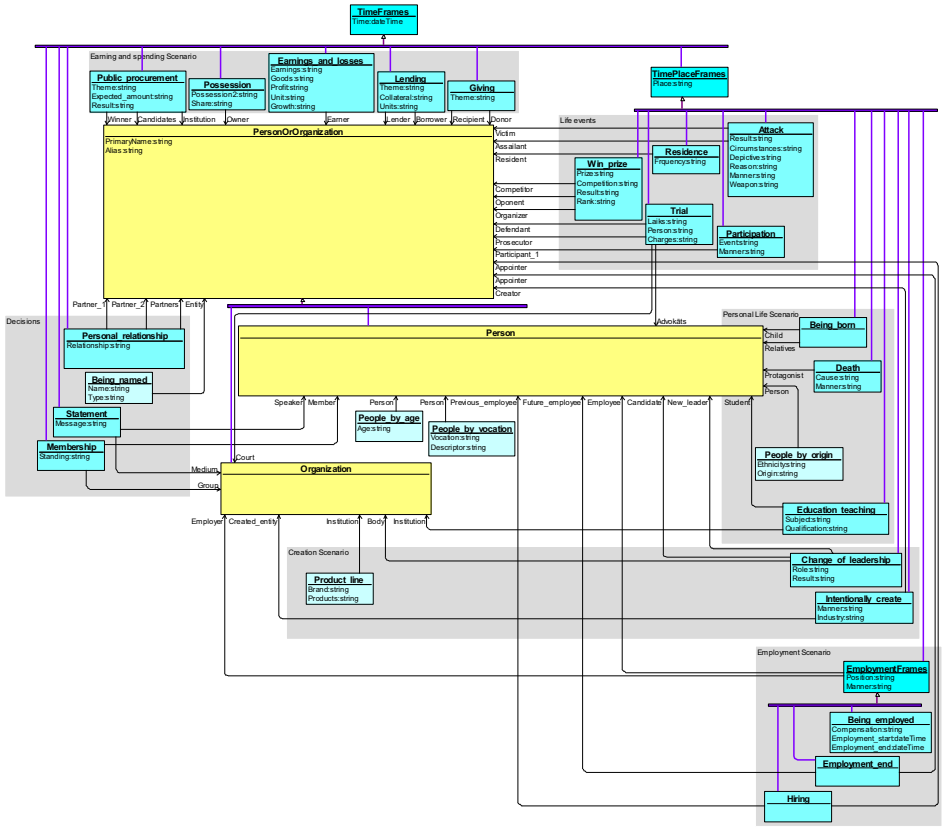

Figure 2. Knowledge representation ontology.

4. Entity Linking

We also implement an entity linking process, matching the entity mentions in each document to an entity knowledge base that is initially seeded with authoritative information on ~25,000 popularly known people and ~35,000 companies, and source data on 'classifiers' such as locations, professions, etc.

Common practice for larger languages such as English is to link entities to the identities listed in large public repositories such as Wikipedia or DBpedia, but for Latvian those resources do not provide a sufficiently high coverage of locally important people and companies. Thus, an internal authoritative list is used, based on data previously aggregated in proprietary systems of LETA. The number of entities rapidly increases by a factor of ten as new, less common entities get added from document analysis – authoritativeness of such entities is expected to be maintained by manually reviewing and correcting newly identified entities with a large number of mentions. The entity knowledge base data is also fed back to the analysis stage in order to improve named entity classification accuracy by using the previously seen entities.

Entity name ambiguity is resolved by a cross-document coreference technique loosely based on the entity linking model used by Wick, Singh et al [9]. It is assumed

that in case of people sharing identical names, the knowledge base would contain a list of those namesakes, and disambiguation can be performed amongst them – and if it is not, then such entities can be flagged for manual review due to conflicting factual data such as different claimed birth years. Three separate signals are used for classification: (a) the components of 'extended names', taking the extended noun phrase parts from document mentions, and appositive items from the known frames such as titles or professions; (b) connected entities – other entities mentioned in the document versus entities sharing a common fact/frame in the knowledge base; and (c) document level context according to a bag of words model, which approximately models document topic –disambiguating an actor and a politician sharing the same name by looking if the document contains words specific to theatre or political reporting.

The similarities between the entity mentioned in newly analyzed document and the candidate 'true' entities are evaluated with a cosine-similarity metric. Similarity is measured against news articles previously marked as mentioning that specific person, or in the minimal data case, against a source 'CV' article from which the relevant properties and context can be extracted.

5. Conclusion

This research shows that information retrieval techniques and natural language processing tools are sufficiently mature that commercially usable information extraction systems for specific domains can be implemented using currently published methodologies and available tools.

The accuracy of automated frame extraction varies depending on specifics of each frame type, but the average accuracy currently is 53% F-score for frame target identification, and 61% for frame element role classification [8]. A prototype of this system has been implemented and at the time of writing this abstract is currently undergoing pilot testing by analyzing news articles mentioning particular individuals, and comparing the automatically extracted data with a human analysis of the same articles. Initial error analysis indicates a strong dependency on the accuracy of initial analysis layers – mistakes in syntactic analysis or named entity recognition cause those elements to be misclassified in the semantic analysis as well.

Surprisingly large part of the system is nearly language independent – while implementing the initial text analysis steps (morphosyntactic structure, entity processing) required a significant amount of language-specific tools, at the stage of semantic role labeling the data is processed in the FrameNet representation, which is domain specific but language neutral. This would enable combining information from sources in multiple languages, if the entity mapping between languages is adequate, joining person and company names in different languages, and also 'classifier' type entities such as professions and family relations.

Adapting the existing system to different domains and languages (assuming that the generic language processing modules are available for the target language) would initially consist of (a) developing a FrameNet model for the desired information mapping and (b) annotating a semantic training corpus according to that model containing approximately 100 examples for each frame.

A specific challenge for Latvian was capturing the notion of semantically similar words in order to reduce the sparsity effect of training data. The published state of art systems for other languages gained an accuracy boost of multiple percentage points by

including WordNet lexical data, which is not available for Latvian. This gap was partially filled by manually developing a number of lists of domain-specific semantic groups of words (synonym sets of targeted verbs, lists of job titles, etc.) and including them as features for the classifiers.

6. Acknowledgements

This work has been supported by the European Social Fund with the project "Support for Doctoral Studies at University of Latvia".

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Center" of EU Structural funds, contract nr. L-KC-11-0003, signed between ICT Competence Centre (http://www.itkc.lv) and Investment and Development Agency of Latvia, Research No. 2.7 "Creation of the New Information Archive Access Product based on Advanced NLP".

References

[1] P. Vossen, G. Rigau, L. Serafini, P. Stouten, F. Irving, W. van Hage, NewsReader: Recording History from Daily News Streams. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation* (2014).

[2] X. Artola, Z. Beloki, A. Soroa, A Stream Computing Approach Towards Scalable NLP. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation* (2014).

[3] R. Steinberger, B. Pouliquen, E. van der Goot, An introduction to the Europe Media Monitor family of applications. *Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop* (2009). Boston, 1–8.

[4] J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, J, Scheffczyk, *FrameNet II: Extended Theory and Practice. Berkeley*, International Computer Science Institute, California, USA, 2010.

[5] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, Abstract Meaning Representation for Sembanking. *Proceedings of Linguistic Annotation Workshop* (2013).

[6] R. Johansson, P. Nugues, LTH: semantic structure extraction using non projective dependency trees. *Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluations* (2007), 227–230.

[7] D. Das, D. Chen, A. F. T. Martins, N. Schneider, N. A. Smith, Frame-Semantic Parsing. *Computational Linguistics,* **40:1** (2014).

[8] G. Barzdins, D. Gosko, L. Rituma, P. Paikens. Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation* (2014).

[9] M. Wick, S. Singh, H. Pandya, A. McCallum, A Joint Model for Discovering and Linking Entities, *Proceedings of the 2013 workshop on Automated knowledge base construction* (2013), 67–72.

[10] X. Han, L. Sun, A generative entity-mention model for linking entities with knowledge base. *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* **1** (2011), 945–954.

[11] V. Stoyanov, J. Mayfield, T. Xu, D. W. Oard, D. Lawrie, T. Oates, T. Finin, A context-aware approach to entity linking. *AKBC-WEKEX '12 Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction* (2012), 62–67.

[12] P. Paikens, L. Rituma, L. Pretkalniņa, Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) NEALT Proceedings Series* **16** (2013), 267–278.

[13] L. Pretkalnina, L. Rituma, Statistical syntactic parsing for Latvian. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) NEALT Proceedings Series* **16** (2013), 279–290.

[14] A. Znotins, P. Paikens, Coreference Resolution for Latvian. *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation* (2014).