Semantic Enrichment Across Language: A Case Study of Czech Bibliographic Databases

Pavel Smrz
Brno University of Technology
Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czechia
smrz@fit.vutbr.cz

Lubomir Otrusina
Brno University of Technology
Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czechia
iotrusina@fit.vutbr.cz

Abstract

This paper deals with semantic enrichment of textual resources by means of automatically generated named entity recognizers-linkers and advanced indexing and searching mechanisms that can be integrated into various information retrieval and information extraction systems. It introduces a new system transforming Wikipedia and other available sources into task-specific knowledge bases and employs contextual information to build state-of-the-art entity disambiguation components. Although some components are language-dependent (for example, that responsible for the morphology analysis or the semantic role identification), they can be easily replaced by existing tools providing specific functions. As a case study, we demonstrate an instantiation of the system for the task of semantic annotation of Czech bibliographic databases in the context of the CPK project. We particularly stress the role of problem-specific knowledge sources that can be easily integrated into our system and play a key role in the success of the tool in real applications.

1 Introduction

Semantic enrichment of textual resources is a well-studied field with applications in many different domains, such as news, market analysis, environmental studies or cultural heritage. Various general-purpose tools have appeared in recent years (some of them are discussed in the next section). Existing systems can often recognize basic entities and provide a link to a knowledge base (KB), but they usually lack additional information on entities that is critical for specialised applications. Indeed, it is not guaranteed that a referred KB entry contains complete and pertinent information, such as task-relevant entity (sub-)type and attributes.

Moreover, it is almost impossible to re-purpose or re-target existing semantic enrichment systems to a new domain or a new context or to adapt and extend it for another language. Clearly, tools for specific domains call for an integration of specific information sources. For example, given a person, a bibliographical system needs information about publications written by him or her, or a list of publications mentioning the person. Although domain-specific knowledge can substantially improve results of entity disambiguation, it is very difficult to make the existing systems use such information.

The ultimate goal of the work reported in this paper is to build a state-of-the-art semantic enrichment system that will be more flexible than existing tools and will need only a minimal effort to adapt to new environments and tasks. The background KB is derived directly from Wikipedia dumps and domain-specific knowledge sources. Hence, it has always up-to-date information and can benefit from recent updates (as opposed to, e. g., DBpedia-based systems relying on the resource updated twice a year which is usually built from Wikipedia versions several months old).

The rest of this paper is structured as follows. The next section surveys existing systems for semantic enrichment and points out differences in the approach they follow. Section 3 deals with the process of KB creation and compares results of the developed method with alternative solutions. Section 4 describes a case study of the semantic enrichment of Czech texts and bibliographic metadata. Last section concludes our work and discusses future directions of our research.

S Bandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 523–532, Kolkata, India. December 2017.

2 Related Work

The need to bridge the semantic gap between the semi-structured "web of documents" and the structured "web of knowledge" (Buitelaar and Cimiano, 2008) has led to the development of various semantic enrichment systems in recent years. Named entity recognition (NER), linking (NEL) and disambiguation (NED) present a key component of the process of semantic enrichment.

Tools such as DBpedia Spotlight (Daiber et al., 2013), Illinois Wikifier (Ratinov et al., 2011), AIDA (Hoffart et al., 2011), or Babelfy (Moro et al., 2014) enable annotating mentions of named entities in a plain text and "anchoring" the annotations in linked open data resources (most frequently in DBpedia/Wikipedia).

State-of-the-art NED methods have to cope with trade-offs among output accuracy, run-time efficiency and scalability. Fast methods, like Illinois Wikifier or DBpedia Spotlight use relatively simple contextual features. These methods perform well on standard texts that deal with prominent entities, but their accuracy is rather low on more complex inputs with highly ambiguous names. On the other hand, sophisticated systems, such as AIDA or Babelfy, rely on rich contextual features, such as key phrases and joint-inference algorithms. Consequently, they tend to be rather slow.

DBpedia Spotlight is a system for automatically annotating text documents with DBpedia[1] URIs. Its disambiguation algorithm is based on the cosine similaritiy and a modification of TF-IDF weights. The main phrase spotting algorithm relies on exact string matching, which uses LingPipes [2] Aho-Corasick implementation.

Illinois Wikifier combines local clues and global coherence of the joint cross-linking assignment by analysing Wikipedia link structure and estimating pairwise article relatedness. It aims at linking all possible concepts to their corresponding Wikipedia pages.

AIDA employs the YAGO knowledge base[3] as an entity catalogue and a source of entity types and semantic relationships among entities. It uses co-occurrence information obtained from large, syntactically parsed corpora as a similarity measure. The AIDA system provides an integrated NED method using popularity, similarity, and graph-based coherence. AIDA-light (Nguyen et al., 2014) is a lightweight version of AIDA. It is a complete system for NED, which is orders of magnitude faster than AIDA while achieving comparable output quality.

Babelfy provides a unified approach to word sense disambiguation and entity linking. It is knowledge-based and exploits semantic relations between word meanings and named entities from BabelNet[4] (Navigli and Ponzetto, 2012) – a multilingual semantic network consisting of more than 13 million concepts and named entities in 271 languages. It is based on a loose identification of candidate meanings (substring matching instead of exact matching) coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations.

Czech named entity recognition has become a well-established field as well. In (Straková et al., 2016), authors present a completely featureless, language-agnostic named entity recognition system. The system uses only surface forms of words, lemmata and tags as its input. Despite that, it surpasses the precision of current state-of-the-art Czech NER systems, which use manually designed rule-based classification features, such as first character capitalization, existence of special characters in the word, or regular expressions designed to reveal particular named entity types. The system is based on artificial neural networks with parametric rectified linear units, word embeddings and character-level embeddings, which do not need manually designed classification features or gazetteers.

Ni and Florian (2017) proposed a new class of approaches that utilize Wikipedia entity type mappings to improve multilingual NER systems. They apply a maximum entropy classifier on English Wikipedia to construct an entity type mapping. To build multilingual mappings, they use the inter-language links of Wikipedia. The system has a fine-grained entity type set containing 51 types (such as person, organization, location, title work, facility, event, date, time…). They use both – structured information, such as title and infobox, and unstructured information, such as text in a Wikipedia page, as features. The classifier trained with all the features identifies the correct

[1] http://dbpedia.org/

[2] http://alias-i.com/lingpipe/

[3] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[4] http://babelnet.org/

category for English with overall F1 score of 90.1. Their approach improved the baseline NER systems for English, Portuguese, Japanese, Spanish, Dutch and German.

Various evaluation campaigns have also recently appeared that compare quality of the NE annotation on collected datasets. Initiatives such as NIST TAC KBP[5] – Knowledge Base Population – Entity Discovery and Linking Track (Ji et al., 2015), NEEL[6] – Named Entity rEcognition and Linking Challenge on Microposts (Rizzo et al., 2015), or ERD[7] – Entity Recognition and Disambiguation Challenge (Carmel et al., 2014) rank participating systems based on their overall performance on collections of specific textual fragments (selected sentences, tweets. . .) that had been manually annotated. As the manual dataset preparation is tedious, the provided training and test data is usually limited to few thousands of entity mentions. Developers of NER tools can measure improvements in annotation quality w.r.t. a particular available dataset or they can use specific benchmarking frameworks such as NERD (Rizzo and Troncy, 2011) or Gerbil (Cornolti et al., 2013), embracing several datasets.

3 Knowledge Base Creation

As mentioned above, we aim at creating a domain-specific knowledge-reference store with the highest possible coverage of entities and specific attributes relevant for a task in hand. Due to its limited applicability across languages and across contexts, we cannot simply apply the approach followed by the DBpedia extraction team (Lehmann et al., 2015). Indeed, specific hand-crafted extraction patterns that are often based on Wikipedia features rare in some languages (e.g., Infoboxes in the case of the Czech Wikipedia) are not easily adaptable to our purposes. On the other hand, we do not want to ignore the Wikipedia structure as a key indicator of entity grouping (and types) and rely solely on pure natural language processing (NLP) of entity definitions. The NLP approach is generally difficult to transfer from one language to another and its success hinges on the discipline authors of Wikipedia articles follow when creating the content.

As opposed to manual approaches, our method employs a straightforward learning technique that capitalizes on the most frequent Wikipedia features in each individual language (most frequently, categories and lists a Wikipedia page is assigned to). The learning process involves two steps of the extraction – identification of entities (more precisely, Wikipedia articles / domain-specific source entries referring to a specific entity) and slot filling of attribute values relevant to a particular task. The overall schema of the Knowledge Base Creation component is presented in Figure 1. Let us focus on the initial step first.

The system enables users to specify a basic set of entity types to be recognized (general ones, such as person, location, event, or domain-specific, for example, visual artists related to a particular place). It then expects a set of examples of (prototypical) representatives of each type. It is also possible to extend the input by negative examples, e.g., entities that are known to be often misrecognized as belonging to a given entity type. The positive examples can be either extracted from other sources such as existing lists of entities of interest, or they can be identified by the user. If the entity type exists in a knowledge source in another language in which entities have been already identified, the system can also take advantage of inter-language links and consider all entities of a given type linked to the language as positive examples for the automatic extraction process. Note that inter-language links are treated as features of the automatic technique described and evaluated below so that one can directly see what value they bring as examples of specific entity types.

Even if the initial set of examples is limited, our method can find a significant portion of all entities of the given type. This can be demonstrated by the results in Table 1 showing the numbers of examples necessary to reach the coverage of 90 % or higher for selected types in English Wikipedia. It is clear that less than 20 examples usually suffice to cover almost all entities of particular types so that even fully manual instantiation of the system does not present a tedious job.

If no negative examples are explicitly provided, the system uses entities of all other types as negative examples for a current type, while considering mutual exclusivity constraints that can be specified by the user (all categories are considered mutually exclusive by default). Depending on the complex-

[5] http://nlp.cs.rpi.edu/kbp/

[6] http://www.lancaster.ac.uk/Microposts2015

[7] http://web-ngram.research.microsoft.com/ERD2014/

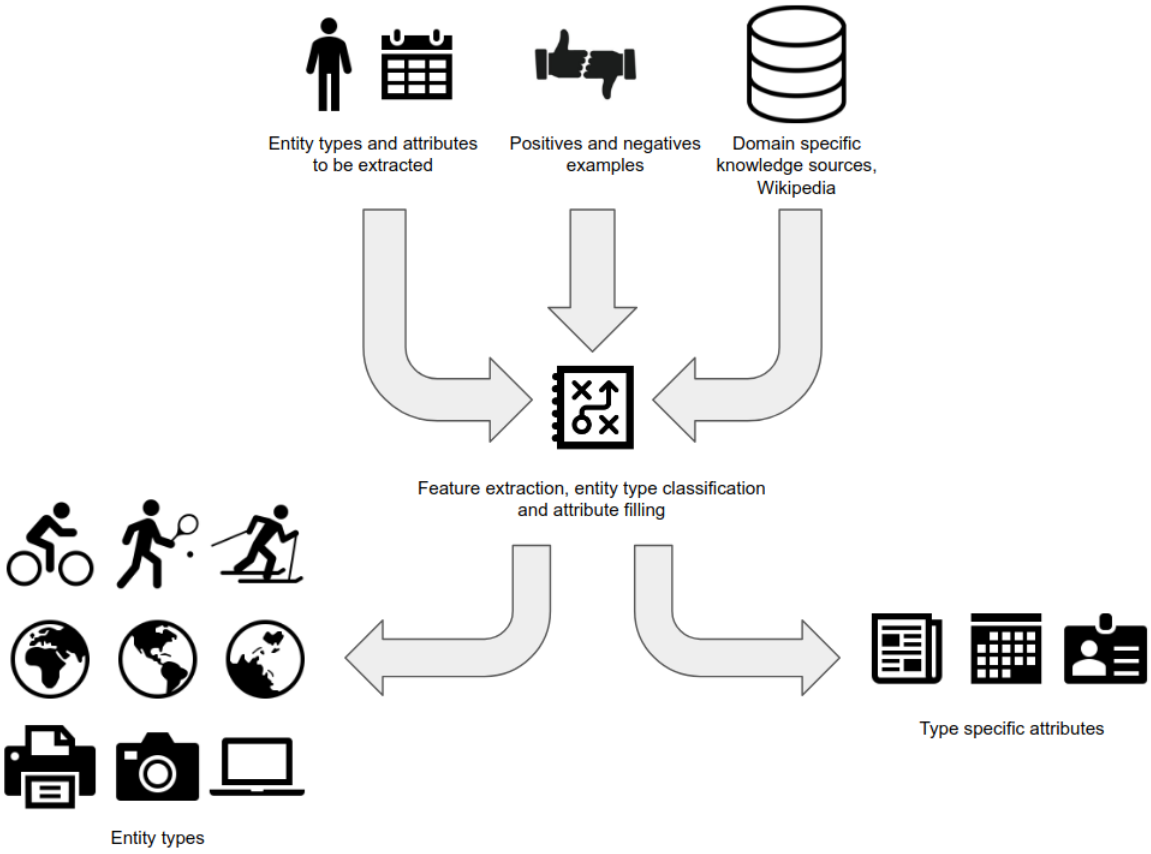

Figure 1: Structure of the KB creation.

Entity type	Min. number of Examples	Coverage
Beverage	10	90.0 %
Musical Work	20	90.3 %
Video Game	15	96.9 %

Table 1: The numbers of examples necessary to reach the coverage of 90 % or higher for selected types in English Wikipedia.

ity of the type, the filtration based on negative examples can bring significant improvements to the automatic extraction.

When dealing with Wikipedia, the learning algorithm explores six types of features:

1. a particular category is assigned to an article;
2. an entity forms a part of a list;
3. a Wikipedia page contains a specific infobox;
4. the key word / phrase of the definition (the first sentence of an article) corresponds to a given list;
5. the first sentence corresponds to a given pattern (for example, an asterisk before a date in parenthesis);
6. the article has an equivalent in a language where the type corresponds to the expected one.

The system generates hypotheses based on provided examples (for example, "all entities in category American Merchants have type Person) and evaluates their relative value. A simple logit model is used to combine the features contributing the most. To find an optimal combination of features, standard measures of the association rule mining are used – support, confidence, lift, and conviction (Bayardo Jr and Agrawal, 1999). Support indicates how frequently an individual feature or a set of features (a feature set) appears in Wikipedia. The confidence measures how often a feature set truly correlates with an entity type, i.e., the proportion of articles corresponding to a feature set F which are known to correspond to type t. The lift is defined as the ratio of the observed support to that expected if F and t were independent. The

conviction can be interpreted as the frequency that the feature set makes an incorrect prediction of the entity type – the ratio of the expected frequency that F occurs with entities of other types than t.

Let symbol $\mapsto$ denote the mapping of a feature set to an entity and E be the set of all entities (or, more precisely, all articles that can deal with an entity). The above-mentioned measures can be formally defined as:

$$\mathrm{supp}(F) = \frac{|\{e \in E; F \mapsto e\}|}{|E|}$$

$$\mathrm{conf}(F \Rightarrow T) = \frac{\mathrm{supp}(F \cup T)}{\mathrm{supp}(F)}$$

$$\mathrm{lift}(F \Rightarrow T) = \frac{\mathrm{supp}(F \cup T)}{\mathrm{supp}(F) \times \mathrm{supp}(T)}$$

$$\mathrm{conv}(F \Rightarrow T) = \frac{1 - \mathrm{supp}(T)}{1 - \mathrm{conf}(F \Rightarrow T)}$$

Table 2 show examples of hypotheses generated by the system for person identification from English Wikipedia along with their support, confidence, lift, and conviction values.

The entity type presents just the initial attribute the automatic system extracts from Wikipedia and/or domain-specific knowledge sources. The user can define a full set of additional attributes corresponding to the entity type that are relevant for a given task. The attributes can reflect the structure of a known template for a given entity type (for example, attributes of a Wikipedia infobox or all potential relations an entity of the type can have in DBpedia). On the other hand, the attribute can be also specific to a given context. For example, our previous work in the cultural heritage domain (Smrz et al., 2013) utilised attributes of art influences and travel experiences of visual artists that could not be directly mapped to a pre-identified relation in DBpedia/Wikipedia infoboxes.

Existing extraction frameworks provide a very limited quality in terms of the completeness of attributes being correctly filled based on information contained in the source. To reach a high coverage and precision, the system applies a learning approach again. Feature detection does not work with the textual content directly. It rather deals with word embeddings (a combination of Word2Vec (Mikolov et al., 2013) and GLOVE (Pennington et al., 2014) vectors) and generalizes the structure of textual fragments corresponding to attribute-filling examples provided by the user.

The resulting knowledge base can be further supplemented by additional information necessary for the actual identification of entities in text, unique identification of entities and their linking to other authoritative knowledge sources, disambiguation information and entity visual representation (if available). As discussed in the following section, entity mentions are matched in text by means of a finite-state technology. The KB should contain all alternative names that can refer to an entity and all forms of names one can expect in the text. For the KB generated from English Wikipedia, we collect and process all alternative names (redirections) and generate known variants of the name forms (for example, shortened or omitted middle names). We also use interlanguage links to record language variants of entity names and consider transcription to characters without diacritics (through Unicode equivalent classes) and transliterations. The generation of name forms in a morphologically rich language (Czech) is discussed in the next section.

We store and later index entity identifiers in their original sources and known interlinks to other LOD (Linked Open Data) resources (for example, Wikipedia URI in both forms with numerical article IDs as well as titles, links to Freebase, DBpedia, etc.). To be able to correctly assign the KB link for an entity with an ambiguous name, we store a vector characterizing words and multiword expressions appearing with an actual meaning of each name. The disambiguation algorithm then reads this data and classifies the entity according to an observed context. If a Wikipedia article referring to an entity contains an image / images, we store this information to be able to represent the entity in a visual way. Figure 2 demonstrates one form of visualization based on such KB.

As already mentioned, the KB resulting from our system covers significantly more entities (correctly assigns the type to more entities) than that of alternative solutions. This is true not only for specific types and less frequent languages such as Czech (see the next section) but also for standard entity types and attributes in English. Table 3 com-

hypotheses	support	confidence	lift	conviction
all entities in category ending with *births* have type Person	23.22	99.51	400.08	15581
all entities whose page contains section *History* don't have type Person	10.17	99.52	132.47	5196
all entities in category ending with *deaths* have type Person	10.79	99.81	401.27	40540
all entities in category starting with *People* have type Person	11.10	99.34	399.40	11536
all entities whose page contains template *coord* don't have type Person	6.48	99.97	133.07	100933

Table 2: Examples of hypotheses for person identification from English Wikipedia.

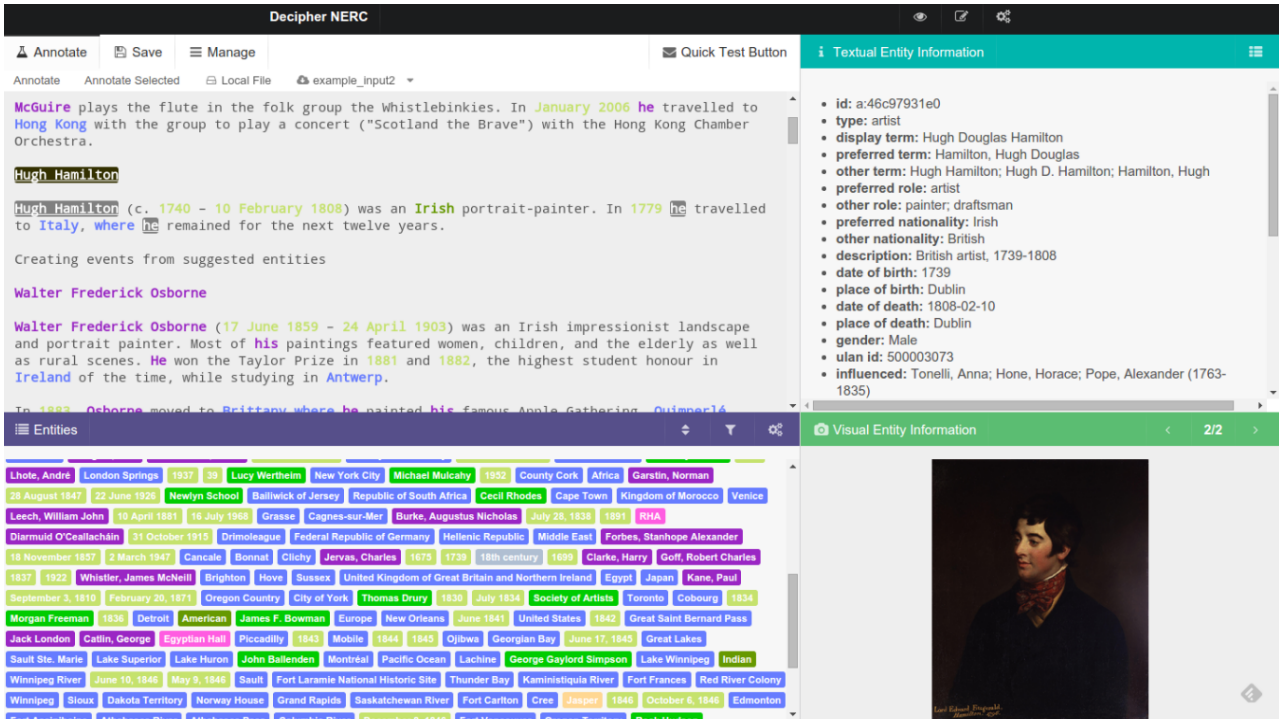

Figure 2: Visualization of an entity.

pares numbers of entities of common general types that can be found using a SPARQL query (often rather a complicated one) in the current DBpedia and the corresponding numbers resulting from our system. It is obvious that if a task requires high precision, it cannot easily rely only on the types identified by DBpedia.

type	number of entities	
	in DBpedia	in our KB
person	1,348,346	1,624,602
place	806,418	738,328
organization	253,215	606,712
video game	20,910	44,862

Table 3: Numbers of entities of common types that can be found using a SPARQL query in the current DBpedia and the corresponding numbers resulting from our system.

4 Semantic Enrichment of Czech Bibliographic Databases

This section presents a case study of the semantic enrichment system applied to Czech bibliographic databases. It forms a part of our work in a large national project – CPK: Using Semantic Technologies to Access Cultural Heritage Through the Central Portal of Czech Libraries. The primary ob-

jective of the project is to research and develop methods leading to a significantly improved access to cultural heritage available in Czech libraries as well as applying results of this research to the creation of the Central portal of Czech libraries (CPK), using advanced semantic technologies. The portal is to be developed in the form of a technological system combining existing and newly developed component applications and a specialized public database (index) of resources integrated into the portal. CPK indexes not only metadata records from contributing institutions, but also full texts of documents digitized by libraries and other information sources.

The fulltexts Czech libraries collect and provide to readers include historical as well as contemporary newspapers and other periodicals, monographs, and various other material of a cultural value. As opposed to bibliographic records associated with monographs that generally contain subject references, including links to authoritative entity knowledge bases (discussed in detail below), the search in the content of newspapers cannot take advantages of the semantically annotated metadata (only a simple fulltext search is supported). That is why our work primary focuses on the available content of periodicals and its automatic semantic enrichment. Note that it is expected that later phases of the project will extend this towards particular categories of monographs (such as historical non-fiction and biographies) and will scale-up the techniques developed for periodicals to a wide range of the cultural heritage artefacts.

Most of the modules integrated into the processing pipeline for the described use case can be easily adapted and used for the same task in any language. We specifically identify the parts that are language-dependent and that need to be replaced by a tool or a resource tailored to the task when the system is to be transferred to another language.

The discussed use case also builds on task-specific knowledge sources. The Czech libraries work with a collection of controlled vocabularies and thesauri, referred to as National Authorities (or simply Authorities), that include named entities (persons, geographical entities, events, etc.). The Authorities are uniquely identified, they provide an official- as well as alternative names, a brief description of the entity, and a link to the Czech Wikipedia (if available). In addition to the information that could be extracted from Wikipedia, the knowledge source can be also used to identify all monographs that deal with particular entities (either in their fulltext form or, at least, as a metadata record listing the title, authors, other subject references, etc.). For the entries corresponding to people, the works authored by a given person can be also identified (currently, only titles and classifications of the work are considered in such cases).

It is critical to recognize that the task-specific knowledge source can bring invaluable quality to the semantic enrichment process and can enable extracting entities and relations not be covered by Wikipedia and other general-purpose resources. This can be demonstrated by the statistics summarized in Table 4. It is obvious that monographs in Czech libraries refer to many entities not addressed by the Czech Wikipedia. To be able to recognize mentions of relevant entities in Czech periodicals and monograph, one cannot simply reckon on Wikipedia only.

A back side of the integration of domain-specific resources lies in a significant increase of ambiguity of names. While there are only 7.1 % of ambiguous person names in the Czech Wikipedia, the knowledge base combining this resource with all Authority files has the ambiguity of 13.3 %. Fortunately, the links to metadata records and fulltexts of monographs dealing with referred Authorities can provide additional learning contexts for the entity disambiguation process. This is also true for entities covered by both Wikipedia and national Authorities so that the additional data actually improves the quality of entity disambiguation models. Indeed, combined Wikipedia pages, referred web links and all available WikiLinks (Otrusina and Smrz, 2016) are not usually a match for an entire book dedicated to a person, location, event, or other type of entities.

Czech is a morphologically-rich (inflectional) language. It is not easy to generate all potential forms in which entity names can appear in text. For example, nouns and adjectives distinguish 7 cases (sometimes with 2 forms per case) in singular and 7 cases in plural and multi-word names of entities come in various forms with complex agreement rules (e.g., genitive phrases with prepositional groups). Moreover, not all parts of multi-word names are capitalized (for example, the Czech Republic would be Česká republika in

type	number of entities in national authorities	covered by Czech Wikipedia
person	629,122	14,758 (2.35 %)
place	29,041	3,386 (11.65 %)

Table 4: Entities from Czech Authorities covered by Czech Wikipedia.

Czech) so that recognition of the boundary of an unknown (new) name can be complicated. The generation of various morphological forms of entity names is clearly language-dependent.

The generation of the name forms was divided into two steps. First, we identified all single-word names and single-string parts of multi-word names that were not covered by an existing morphological database/analyser. We employed a statistical learning method that considers various morphological features (esp. word endings and frequency of similar word paradigms) and estimates the most probable morphological paradigm according to which the word would decline. We then generated all potential wordforms and corresponding morphological tags for single-word names and prepared the same for filtering multi-word combinations by means of group agreement rules.

The second step dealt with multi-word entity names only. Known names served as training examples for statistical name-structure prediction method. It showed to be beneficial to recognize not only the structures necessary for name-form generation, but also to estimate precise meaning of the name components in this phase. The algorithm was trained to distinguish the most probable given names from surnames, titles and name specifiers (such as the younger), numerical components of the names (WWI), etc. This was critical for organizing the names into hierarchies as well as for generating variants with name initials and shortened forms, acronym matching, etc.

The process of semantic enrichment of the Czech text is realized in the form of a pipeline combining various text- and language analysing tools and task-tailored classifiers. The input can be provided as a plain text (for example, the content of historical periodicals obtained by means of OCR) or in the HTML form. We employ CLD3 [8] for language identification (currently filtering out all non-Czech documents) and JustText (Pomikálek, 2013) to identify and eliminate potential boilerplate in HTML Depending on the further planned processing, the text can be tokenized and verticalized in this phase and task-relevant links from the HTML can be stored to be considered later too. To enable linguistically-oriented search in the indexed material, the text can be also lemmatized, PoS tagged and parsed. The relevant tools can be easily plugged-in as well as replaced if necessary. We take advantage of UDpipe (Straka and Straková, 2017) – a linguistic processing pipeline providing results in various languages.

The actual entity recognition and disambiguation is realized by the SEC component developed by our team (Dytrych and Smrz, 2016). The tool is publicly available[9] and can be easily instantiated for other contexts and new languages. It takes the knowledge base (in the form described the previous section) and extracts all potential entity names (a primary reference name, all alternative names, morphological forms, and generated variants). A minimum finite-state automaton is constructed from all the name strings and corresponding indices in the knowledge base by means of an algorithm described in (Daciuk et al., 2000). The SEC recognizer is thus able to associate a textual fragment that corresponds (can correspond) to a name with all potential KB entries it can refer to. Using the information stored in the KB (for example, statistics on Wikipedia article popularity) it can also provide prior probabilities of the potential entity linking. The SEC can also plug-in other entity disambiguation modules that take advantage of additional contextual information stored in the KB. The current version for Czech combines a rule-based disambiguation module implementing strict application-specific restrictions (for example, preferring entities covered by national Authority files) and a learning-based disambiguator trained on all the relevant material available in the CPK project. To enable users to easily extract semantic relations and to search the resource semantically, the SEC engine can also incorporate coreference resolution tools, semantic role labellers and

[8] https://github.com/google/cld3

[9] http://sec.fit.vutbr.cz/sec/

other language processing components.

5 Conclusions and Future Directions

The semantic enrichment system introduced in this paper stresses the easiness of system customization and transfer to other languages. We showed that it is worth to pay attention to preparation of the knowledge base component generated from Wikipedia and domain-specific resources. The presented learning-based extraction system leads to better coverage and generates data directly applicable in the entity recognition and linking task. Moreover, entity types and attributes are not "hardwired", they are fully defined by application needs. Results of the case study on semantic enrichment of Czech texts collected in the CPK project further demonstrate that existing knowledge-reference systems bring a significant value to the semantic enrichment process and that a system based just on general-purpose knowledge sources would not satisfy the need of specific applications.

There are several directions of research we will explore in our future work and apply in the CPK project. We are going to extend interlinks with DBpedia, KBpedia, and other resources and take advantages of established hierarchical structures (ontologies) to initialize the set of prototypical attributes in order to simplify the task of type-specific attribute definition. We will also integrate an advanced classifier of specific entity types and attributes that are not covered by the KB. Less known people, places, events and other entities are often introduced in newspaper texts (usually the first time they appear) so that the system could suggest extensions to the national Authority database adding the most cited new entities. As for the search interface, a lot needs to be done to make the system more intuitive and accessible to laymen.

Acknowledgments

The work reported in this paper has been supported by the Ministry of Culture of the Czech Republic, programme NAKI II, project DG16P02R006 CPK: Using Semantic Technologies to Access Cultural Heritage Through the Central Portal of Czech Libraries.

References

Roberto J Bayardo Jr and Rakesh Agrawal. 1999. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154. ACM.

Paul Buitelaar and Philipp Cimiano. 2008. *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. IOS Press.

David Carmel, Ming-Wei Chang, Evgeniy Gabrilovich, Bo-June Paul Hsu, and Kuansan Wang. 2014. ERD'14: Entity recognition and disambiguation challenge. In *ACM SIGIR Forum*, volume 48, pages 63–77. ACM.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 249–260, New York, NY, USA. ACM.

Jan Daciuk, Stoyan Mihov, Bruce W Watson, and Richard E Watson. 2000. Incremental construction of minimal acyclic finite-state automata. *Computational linguistics*, 26(1):3–16.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Jaroslav Dytrych and Pavel Smrz. 2016. Interaction patterns in computer-assisted semantic annotation of text - an empirical evaluation. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,*, pages 74–84. INSTICC, SciTePress.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*, pages 782–792.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. Aida-light: High-throughput named-entity disambiguation. In *LDOW*.

Jian Ni and Radu Florian. 2017. Improving multilingual named entity recognition with wikipedia entity type mapping. *arXiv preprint arXiv:1707.02459*.

Lubomir Otrusina and Pavel Smrz. 2016. Wtf-lod - a new resource for large-scale ner evaluation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

J Pomikálek. 2013. justext: Heuristic based boilerplate removal tool. *Available: Google code, online http://code. google. com/p/justext*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics.

Giuseppe Rizzo and Raphaël Troncy. 2011. Nerd: evaluating named entity recognition tools in the web of data.

Giuseppe Rizzo, Amparo E Cano, Bianca Pereira, and Andrea Varga. 2015. #microposts2015 named entity recognition & linking challenge. In *5th International Workshop on Making Sense of Microposts*.

Pavel Smrz, Lubomir Otrusina, Jan Kouril, and Jaroslav Dytrych. 2013. Decipher deliverable D4.3.1: Semantic Annotator. Technical report, Brno University of Technology, Faculty of Information Technology.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2016. Neural networks for featureless named entity recognition in czech. In *International Conference on Text, Speech, and Dialogue*, pages 173–181. Springer.