An Interoperable Platform for Multi-Grain Text Annotation

Svetlana Sheremetyeva

South Ural State University, 76 Pr. Lenina, Chelyabinsk, 454080, Russia

Abstract

In this paper, we describe an interoperable platform for creating annotated corpora in different languages and domains. It focuses on two most widely used for practical information processing tasks levels of linguistic annotations, - morphological and conceptual, that can be performed separately or combined. The platform consists of two main modules, - a program shell and a knowledge base. The program shell is universal and features flexible settings that ensure its adaptation to multilingual corpora of various domains and different levels of annotation. It is provided with several interfaces for knowledge acquisition and annotation control. The annotation platform knowledge base includes language-independent and language-dependent linguistic information. The language-independent information is presented by multilingual domain ontology, while the core of the language-dependent component of the platform knowledge base includes unilingual onto-lexicons. The annotation process consists in the practical realization of ontological analysis. In performing the annotation task, the NLP techniques are used to automatically support, rather than completely replace human judgment. The platform is multifunctional, and in addition to corpora annotation, it can directly be used for different types of theoretical linguistic research, e.g., terminology analysis, cross-linguistic comparative studies, etc. The paper covers both, the platform design and its application in the frame of a real project on the conceptual annotation of the "Terrorism" domain corpora in the Russian, English and French languages.

Keywords

Annotation platform, interoperability, domain ontology

1. Introduction

Corpora annotations are a prerequisite for any succession of text processing steps and its accuracy to a large extent defines the quality of the final NLP output. It is therefore the focus of many international theoretical and applied linguistic studies. While many practical texts processing tasks nowadays rely on morphological labelling, conceptual annotation is becoming increasingly used as explicit semantics is starting to play a more prominent role in computer technologies targeted to intelligent processing of unstructured information (automatic classification, intelligent content and trend analyzes, machine learning, machine translation, etc.) [1]. By conceptual annotation (which in many practical projects is called “semantic”) we understand that type of semantic annotation, which is developed for solving specific information tasks within a particular domain, and use the term to distinguish this particular type of annotation from the high level semantic mark-up such as “human”, “animated”, etc. For example, in the “Terrorism” domain the English lexeme “car” will be conceptually annotated as “means of attack”, rather than “concrete”, “non-animated”, etc. We also believe that given the ambiguity of natural language on all levels, combining different types of annotations, e.g. morphological-syntactic and conceptual might provide a feature space that would enhance the chances to resolve annotation ambiguity.

Information processing projects that strive for high quality results require annotating comprehensive corpora, which with any level of tags, let alone conceptual, as a starting point of research and development is nowadays mostly done manually and on its own is a hard, costly and

IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: sheremetevaso@susu.ru
ORCID: 0000-0003-1245-4213

time-consuming task. Taking advantage of pre-developed resources that could allow skipping the annotation stage is quite problematic. Annotated corpora are quite sparse and often cannot be accessed at all, because the developers restrict or completely forbid their public use. In addition, the volume and construction principles of most existing annotated resources are non-standardized and are tuned to only a limited number of domains and information processing tasks. The situation puts in focus the issues of developing automated annotation tools and their interoperability to save development effort.

This paper attempts just that and presents an automated interoperable platform for creating multi-grain annotations of corpora in different languages and domains. The platform is ontology-based and is supported by the NLP technology that complements human annotation effort. The tool is multifunctional. In addition to automated corpora annotation, it can directly be used for different types of theoretical linguistic research, e.g., terminology and corpora analysis, cross-linguistic comparative studies, etc. The description covers both, the platform design and its application in the frame of a real project on the conceptual annotation of the "Terrorism" domain e-news in the English, Russian and French languages.

The paper is structured as follows. Section 2 overviews the related work. Section 3 describes the platform design. Section 4 is devoted to a case-study, in the frame of which the platform development and its use is described as applied to the multilingual corpora of the "Terrorism" domain in English, Russian and French. We conclude with the summary and future work.

2. Related work

While all annotated corpora created to date necessarily contain morphological markup, since the problem of automatic (or automated) morphological analysis for a large number of languages has now been largely solved, the need to speed up and save human effort in corpora annotation for intelligent text processing applications prompted studies specially devoted to the development of automated concept annotation tools. Some attempts are made to apply unsupervised approaches and completely exclude human labor [2]. However, most popular are semi-automatic approaches that rely on NLP techniques [3], document structure analysis [4] or learning that requires training sets or supervision [5]. Some works to automate annotation rely on information extraction [6, 7]. Most modern semi-automatic annotation tools are based on ontologies where the annotation procedure is performed by the technique of ontological analysis that results in the identification concept instances from the ontology in texts [8]. Notwithstanding whether ontology-based annotation is done manually or involves automation, it has a very serious limitation, - the availability of an appropriate pre-defined and well-established ontologies. Though quite a number of ontological libraries are now publicly available, their suitability for every particular R&D project involving ontology-based conceptual annotation is, as a rule, problematic. Most works on ontology-based annotation therefore assume the availability of an already existing ontology [9] or include the creation of an ontological resource as part of annotation problem solution. Ontologies are mostly created for conceptual annotation of domain corpora in one (often, English) language and are tuned to specific information processing tasks, - medical record analysis [10], personalized filtration of eNews [11], “Terrorism” domain content analysis [12]. Much less research can so far be found on the ontology-based annotation of corpora in other languages. For example, in [13] research on the semantic (in fact, conceptual) annotation of the Russian e-service domain corpus is described as presented in e-news, the system presented in [14] focus on the conceptual annotation of the French corpus. Most often, the methodologies for the ontology based annotation include a combination of automated technics and manual tagging (see e.g., the works cited above).

Given the amount of effort and time needed to construct ontologies for language-specific corpora processing, multilingual ontologies that could be interoperable cross-linguistically got in the circle of research interest. There is no consensus on how to understand multilingualism in ontologies. Within one approach, ontological multilingualism is treated as understandability (or adaptation) of the ontological labels for the users who speak different national languages. In another approach, ontology is taken to be multilingual, if it can be applied to processing texts in different languages no matter

what language was used for concept labels. These interpretations of ontological multilingualism directly rely on ontology definition as either a language-independent or language-dependent resource.

Language-dependent ontologies, a well-known example of which is the famous WordNet [16], are thesaurus-like structures defined by the properties of a particular language. Transition to multilingualism there is treated as the localization of ontological concept labels. The localization itself can be approached in different ways, as a) linking the word senses of different national languages to ontological concepts by means of a specially developed model [16], b) translation of the ontological concept labels from one language into another [17] and c) manual annotation of ontological concepts with labels worded in different languages [18]. Among other ontology-related works in the frame of interoperability are, for example, a research devoted to the creation of universal tools for semi-automatic building of unilingual ontologies [19] and the studies to suggest interoperable methodologies for cross-referencing the data and meta-data of unilingual ontologies [20].

Language-independent ontologies, such as Mikrokosmos [21], SUMO [22] and BFO [23], per definition allow multilingualism in the sense of the capability to process texts in different languages, cross-linguistic conceptual annotation included, which is provided by building lexicons of specific languages and mapping them into the concepts of one and the same multilingual ontology.

One of the annotation challenges, which is discussed in the literature, is a way to find the best set of tags for different levels of tagging from morphological tags up to conceptual labels. The main thing here is to decide on the amount of information coded in a single tag, and on the size of the tagset. Though most of the discussions on the tag subject concern morphological and syntactic tagging, the main ideas of such discussions are worth to be taken in consideration for conceptual tagging as well. For example, in [24], the external and internal criteria in a tagset design are suggested. The external criterion demands the tags to be able to code the distinctions in the linguistic features that are required by the processing task. The internal tag design criterion concerns making the tagging process as precise as possible. It is believed that a smaller and simpler tagset should improve the accuracy of tagging, while a large number of tags causes problems for creating reliable taggers. However, larger amount of information included in the tagset may help tag ambiguity resolution. In [25], it is claimed that tagging precision (or accuracy) depends crucially on using a wide range of linguistic features including lexical ones. There is thus the eternal trade-off: tag coverage versus tag precision. Another way to significantly reduce the number of tags and nevertheless take advantage of additional linguistic knowledge for raising annotation accuracy is the use of supertags. In general, a supertag can code a wide range of features (morphological, syntactic, semantic and conceptual thus providing for significant gain in tagger performance [26]. Certain attempts have been made to develop multilingually universal tagsets. Thus, the results of the experiments carried out on different language families (Roman vs. Slavic) are reported and the most challenging linguistic phenomena for the task are defined. Another suggestion is to use a coarse tagset consisting of twelve cross-language lexical categories [28]. In the frame of the MULTEXT-East (MTE) project, an attempt is made to standardize the tagset for a range of Slavic languages, such as Romanian, Croatian, Slovenian, Czeck and, currently, Macedonian and Russian [29]. However, many studies aimed at developing real world applications point out that general-text tagsets usually fail on domain specific texts, and therefore, tagsets should be domain- and application-specific [30].

As noted in [31], current applications using concept tags (or codes) show three different approaches for concept tag definition, - conventional, directed and summative that mainly differ in the tag origin. In the conventional approach, conceptual tagging categories are derived directly from the text data. The directed approach for the initial set of concept tags relies on a theory or relevant research findings. Concept tags within the summative approach coincide with preliminary extracted text keywords. Most often, conceptual tag set design concerns the ontology size and granularity. In [32] the ontological granularity is treated in terms of ontological levels, while the reduction of the number of concept tags is suggested by using specific levels of the so-called multilevel ontologies which would allow meeting the interoperability demand with multi-layer corpus annotation. One more way to save annotation effort concerns the development of cross-platform interoperability for collaboration in automated text annotation [33]. However, in spite of the development of increasingly convivial and hardware-independent annotation tools, the need to create intuitive, user-friendly interfaces, which can make the annotation tools more accessible to users without special technical skills (for example, linguists or domain experts) is more and more emphasized [34, 35].

3. Design

3.1. Overview

Our research and development effort is defined by the intersection of the following criteria: (i) domain and cross-language interoperability (ii) increase of annotation quality, (iii) automation, (iv) user-friendliness for linguists and domain expert’s with-out special technical skills, (v) annotation multi-granularity from morphology up to semantic and conceptual mark-up.

The requirements of annotation interoperability and multi-granularity were answered by defining the annotation methodology as the practical realization of ontological analysis based on a domain-specific multilingual ontology, a universal program shell and a reusable tagset. In defining our tagset features we aimed at providing a) balance between the features’ annotation relevancy and realistic expectations to detect them automatically, b) possibility to disambiguate the tags using both statistical measures and local context linguistic rules as the quality of annotations depends upon the judicious application of NLP technology, and c) possibility to share the tagset between languages within a particular domain. The integration of these methodological and technological solutions determined the architecture of the annotation platform, which consists of two main components - a knowledge base and a program shell. The overall architecture of the annotation platform is shown in Figure 1.

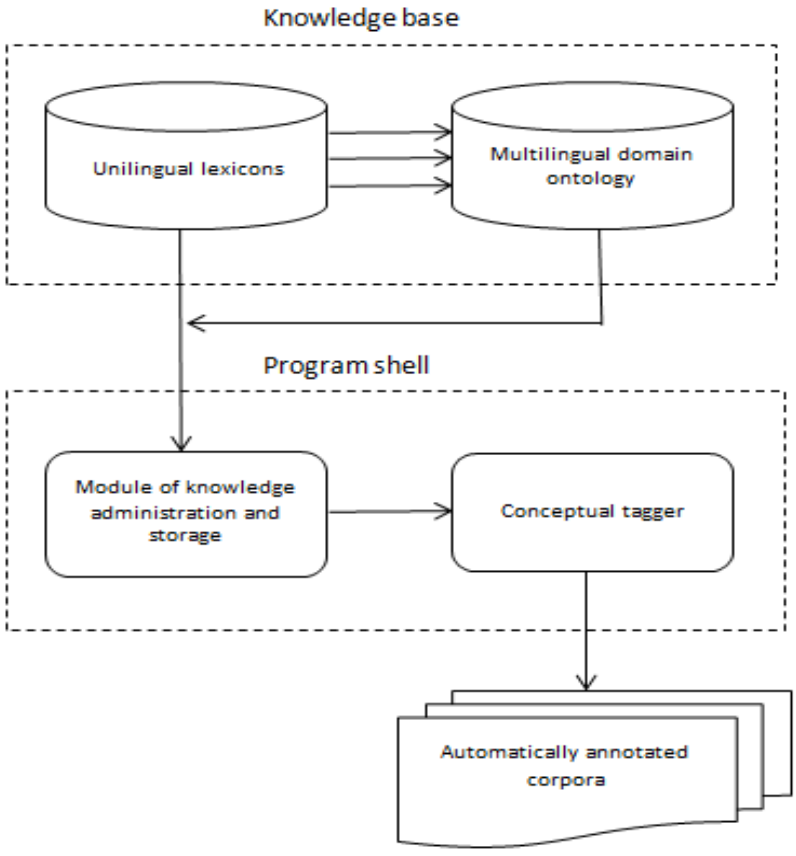

Figure 1: The architecture of the interoperable multi-grain annotation platform

3.2. The knowledge base

The annotation platform knowledge base has the following main components:

- language-independent semantic (conceptual) knowledge of a particular domain presented in the domain ontology;

- language-dependent linguistic knowledge of the domain in question that includes domain-relevant unilingual lexicons of one- and multicomponent units with assigned parts-of-speech and other morphological features relevant for each language;
- linking knowledge on mapping the domain-relevant lexical units into the ontology concepts.

The ontology as the core of the platform knowledge is built based on the following methodological assumptions:

- Ontology is a language-independent resource and serves intermediary between unilingual lexicons.
- Domain ontology is integral part of upper-level ontology, Mikrokosmos [21] in our case.
- The acquisition of the domain ontological knowledge is data-driven based on multilingual comparable domain corpora using mixed (top-down/bottom-up) acquisition techniques.

Building the knowledge base includes extraction of domain-relevant lexemes from training multilingual corpora followed by grouping the resulted sets into semantic (conceptual) categories according to the sense closeness within the one language, and across languages. Thus defined semantic categories are taken to be the seed concepts of the domain ontology and following the Mikrokosmos structure are divided into interrelated classes of the OBJECTS, EVENTS, and PROPERTIES top concepts. The concept labels are worded in English, while the concept meanings are specified by concept definitions. The unilingual lists of domain-related lexemes grouped into conceptual categories are further called onto-lexicons and cover the linking knowledge.

The interoperable annotation platform program shell consists of two main blocks: a knowledge administration and storage module and a tagger (see Fig. 1).

3.3. The program main modules

The main modules of the annotation platform program are a knowledge administration and storage module, further TransDict, and a tagger that are two updated and reused components of the earlier developed text processing platform described in [37] that to a large extent meets our design requirements and allowed us reducing the development effort.

TransDict is structured as a set of unilingual lexicons with cross-referenced entries of translation equivalents. The linguistic information associated with every unilingual entry is formalized as a tree of features:

[semantic class/concept [language [part-of-speech [other morphology [tag]]]]]

The morphological zone of the module entries contains a full wordform paradigm of a unilingual lexeme, each associated with a supertag that codes conceptual and morphological knowledge. The entry is meant for one sense of a lexical unit. TransDict has a powerful environment for the automated acquisition and administrations of multilingual lexical and ontological knowledge by means of a user interface, which visualizers the platform knowledge (Fig.1) and gives access to the following built-in supporting tools:

Configuration block that creates and edits the TransDict feature settings such as semantic classes (concepts), languages, parts of speech, word forms and their tags; any change in the settings will automatically propagates to all the entries in a corresponding language.

Defaulter that automatically assigns entry structures and some of the feature values to new entries according to the user-set parameters and values; for example, all semantic classes and some of the knowledge of the English entry are automatically ported to a lexicon in another language, when added; the knowledge can be edited.

Data importer/merger that imports wordlists and/or feature values from external files and applications both, in batch mode and individually.

Data exporter that exports wordlists and/or feature values from TransDict to external files and applications.

Copy-entry module that copies all, or individual fields of one entry into another

Morphological generator that automatically generates wordforms for a given word and fills the morphological fields of the entry automatically assigning the tags specified in the configuration settings.

Content and format checker, which reveals incomplete and/or ill formatted entries.

Look-up tool that performs a wild card search on one or any combination of specified parameters (letters, language, semantic (conceptual) classes and part-of-speech; it is also possible to filter the whole sets of TransDict entries according to a specified lists of lexemes, incompletely filled entries, entries of repeated tokens, etc. The use of the *Look-up tool* allows identification of knowledge gaps and gives a lot of opportunities for analyzing the qualitative and quantitative linguistic characteristics of the domains, which are either language specific, or hold across languages, and can be used to develop metrics for resolving tag ambiguity (unavoidable in annotations) or for contrastive linguistic research.

To provide for a collaborative setup for sharing knowledge acquisition tasks, TransDict is programmed in two versions: the MASTER version with the full range of built-in tools activated and the LIGHT version, - an empty TransDict program shell configured as MASTER but with the Configuration block disabled for consistency of the acquired knowledge. Acquirers can individually fill LIGHTs with new lexical-ontological knowledge, which is then imported into MASTER on a regular basis.

The platform tagger gets a "raw" text as input and outputs its annotated version at a specified level, - with either conceptual tags only or supertags. The main blocks of the tagger program are as follows:

Configuration block configures the tagger to a specific language and markup level.

Lexicon look-up module tags text with TransDict (super) tags of a selected level

Data importer imports texts from external files and from TransDict knowledge.

Data exporter has two functions: it exports the annotated text to external files and interactively exports lexical units tagged as “unknown” to the TransDict knowledge.

Control interfaces for visualizing tagger output to control the annotation quality.

Disambiguation rules interpreter integrates the rule-based NLP techniques into the annotation process; the interpreter has a user-friendly interface for writing tag disambiguation rules in the simple IF-THEN-ELSE-ENDIF formalism that does not require programmer’s skills. The rules are written over the lexical knowledge and TransDict tagset and, when saved, are automatically compiled into the program. The tagger disambiguation interpreter is fully functional and with a good rule coverage insures the high quality of annotation. Of course, though the interpreter has a lot of effort saving functionalities, the inherent problem of all rule-based NLP techniques (knowledge bottle-neck) cannot be avoided. The interpreter module is therefore made optional and its use depends on the user's willingness to invest a sufficient amount of effort in the disambiguation rule acquisition.

4. Case study: the “Terrorism” domain annotation platform

4.1. Knowledge handling

To be applied in practice, the annotation platform program shell should be filled with domain knowledge along the lines given in Section 2.1. We further describe this process as done in the frame of the real on-going project on content analysis of the “Terrorism” domain e-news in the English, Russian and French languages. The major project task requires the conceptual level of annotation as a must prerequisite.

The main parts of the platform knowledge base, - the “Terrorism” domain multilingual ontology and unilingual English, Russian and French onto-lexicons were built in parallel on the data of three comparable corpora of e-news on terrorist acts of 500,000 words each. The knowledge acquisition details are described in [38]. We here concentrate on its presentation and handling in the TransDict program. A fragment of the TransDict main interface is shown in Fig.2.

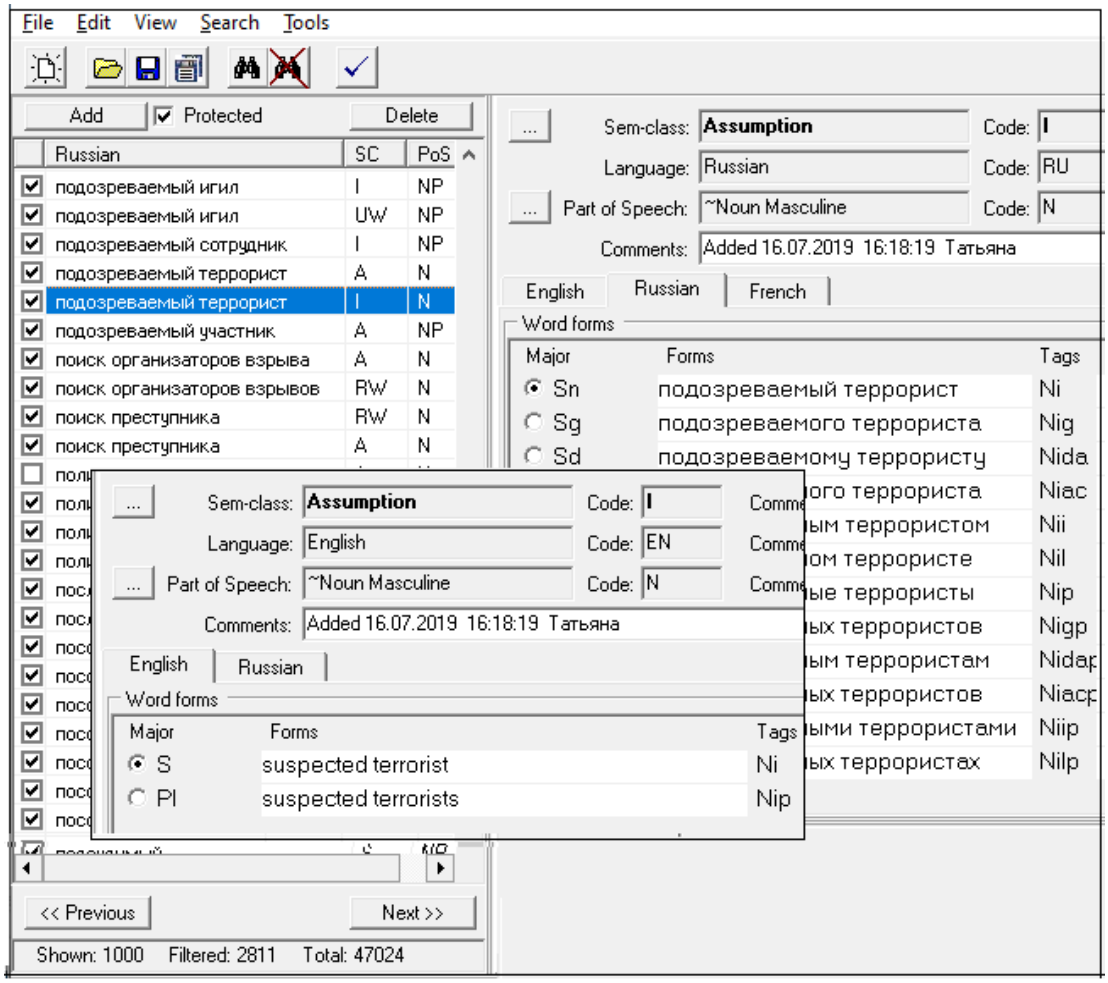

Figure 2: A fragment of the main TransDict interface opened at the Russian onto-lexicon page

In Fig.2, the screenshot of the main TransDict interface displays the entry of the highlighted lexeme. In the center, the pop-up window of its English equivalent entry is shown as called by clicking on the "English" bookmark. The interface buttons are self-explanatory. All fields are interactive and can be edited. On the left pane (from left to right), shown are the interactive list of the Russian onto-lexicon units, corresponding ontology concept codes (SC) and parts-of speech (PoS). Every entry contains a lexeme linked to one ontological concept. In case a lexeme can be mapped into different ontological concepts it appears in different TransDict entries (one per each conceptual meaning). That explains the lexical duplications in the list.

The content of a lexical entry opens on clicks on the lexeme and the bookmark of the language of interest. The knowledge put in the highlighted entry appears on the right pain. The concept, language and part-of-speech with their codes are located on the top of the wright pane, under which the morphological zone containing a full paradigm of a lexeme wordforms with supertags is shown. The TransDict supertags and parts-of-speech are the unified sets of the combinations of task-tuned linguistic features of the Russian, English and French languages; the number of fields in the morphological zone is different and defined according to the grammas of corresponding languages. The new knowledge can be exported to TransDict in a batch mode or individually as follows. A click on the "Add" button over the lexeme list calls the pop-up interactive menu of concepts; the selection of a concept opens the part-of speech menu (see Fig.3), which, in turn opens a new TransDict entry with the selected structure and all the knowledge but the morphological paradigm filled out. The acquirer needs to fill only one wordform in the paradigm field, the rest word-forms will be generated automatically. The content of every entry zone is editable and can be copied from one entry to another. All settings are configurable; the setting changes automatically propagate to the lexical entries.

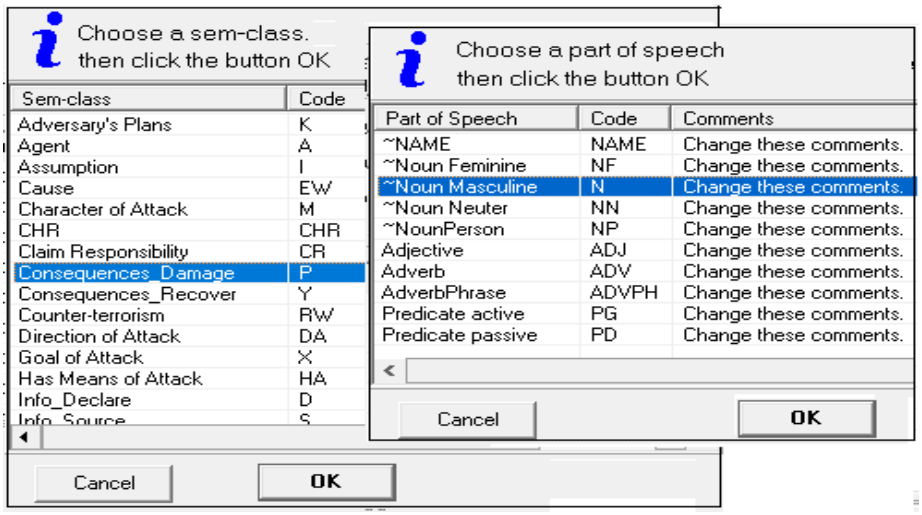

Figure 3: Pop-up windows for assigning a new lexeme its linguistic features

The “Terrorism” domain corpora-based lexemes exported to the TransDict unilingual lexicons is aligned as translation equivalents; the translation gaps are filled out by the acquirers. This augmented onto-lexicons and made the platform useful for machine translation-related tasks as well. The number of aligned lexicon entries is thus the same but the number of unique unilingual lexemes differs due to the different levels of synonymy in each language. The explicit list of lexemes’ paradigms in the TransDict entries allows skipping many analysis problems and annotating the input text by a simple look-up in the TransDict morphological zones. The look-up procedure goes from left to right, longer units first. The results of such look-up can be displayed in the tagger interface on the level of concept tags only (see Fig.4) or on the level of supertags, if necessary.

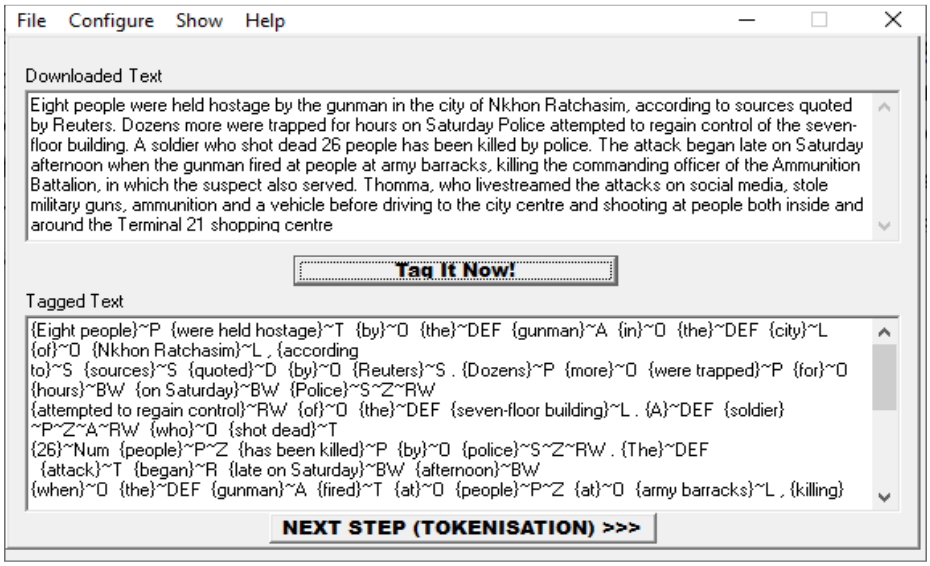

Figure 4: The tagger interface with the concept-only level of annotation after TransDict look-up

4.2. The annotation platform as a research tool

The developed annotation platform due to its advanced search functions accessible through the TransDict main interface can also be used as a research tool. We did just that in an attempt to find quantitative disambiguation metrics that could complement or even substitute the disambiguation

rules. As a first step on this way, we sorted out all lexemes that were linked to multiple ontology concepts and thus lead to conceptual multi-tags after the TransDict look-up. Analysis of both, the sorted out lists and the domain corpora showed that multi-tags are caused by two different phenomena, that of lexical conceptual ambiguity and that of conceptual syncretism. The unilingual lexemes are conceptually ambiguous, if in the domain corpora, they can function in different mutually exclusive conceptual meanings, like, for example, the English word “car” and its Russian and French equivalents “автомобиль” and “voiture”, correspondingly (annotated with the multi-tag ~P~C) :

CONSEQUENCES-DAMAGE (P): The terrorist attack damaged about 50 cars. / В результате атаки террориста повреждено около 50 автомобилей/ L'attaque terroriste a endommagé environ 50 voitures.

MEANS OF ATTACK (C): A car hit people on Westminster Bridge. / На Вестминстерском мосту автомобиль наехал на людей/ Une voiture a heurté des gens sur le pont de Westminster.

The unilingual lexemes are conceptually sincretical, if they have several conceptual meanings that do not contradict each other. Most often, but not exclusively, conceptual syncretism was detected in multicomponent domain-relevant lexemes. For example, in the English noun phrase "airport shooting suspect" the word "shooting" contains information about the type of attack, the word "airport" indicates the place where the attack occurred, the word "suspect" has two conceptual meanings at once - "assumption" and " perpetrator of a terrorist act ”. Therefore, after the tagger look-up this lexeme will be conceptually annotated as {airport shooting suspect} ~T~L~I~A.

In the multi-tag syncretism case no ambiguity resolution is required as the meanings of the individual conceptual tags in a multi-tag are complimentary. On the contrary, multi-tags that are caused by conceptual ambiguity need to be disambiguated. We tried to answer the question whether it is possible to automatically identify syncretical multi-tags to exclude them from the disambiguation procedure.

To reduce the volume of annotator tasks, we conducted the research on relatively small portions of the unilingual e-news corpora of 35,000 wordforms each, which were automatically annotated by the tagger TransDict look-up and manually post-edited to the gold standard. We then calculated the frequencies of the multi-tags, which “survived” the postediting and thus were sincretical per definition. The threshold for cutting the top frequency list of the syncretical multi-tags to be excluded from the disambiguation procedure can be defined empirically. We currently experimented with the 10 top sincretical multi-tags in every language. We further introduced a heuristic *concept usage relevancy* (CUR) measure. The heuristics is: the higher the concept CUR value, the more prioritized its tag can be in the set of the other tags assigned to the same lexical unit. The empirical formula we use at the current stage of research to calculate the CUR value is:

CUR = (RCF*w1+Cf*w2) / (w1+w2), were

RCF is the ratio of concept fillers; it accounts for the variety of the lexical units mapped into a concept and is calculated as

RCF = n/N, where

n is the number of unique (different) unilingual corpus lexical units mapped into a particular concept in the corpus and N is the total number of ontology-mapped lexemes in the corpus;

Cf is the concept frequency index calculated as

Cf = (Cfs +Cfm) / F, where

Cfs is the frequency of the concept in the corpus as a single tag, Cfm is the frequency of the concept in the corpus as a component of a multi-tag; F is the total number of conceptual tags (single and multiple) in the corpus; w1 and w2 are arbitrary weights; we so far experimented with w1= 10 and w2=1. The denominator (w1+w2) in the CUR formula is used to normalize the CUR value to the common percentage scale.

The suggested disambiguation measures are supposed to be crosslinguistically universal, while their values are obviously language-dependent. The scope of this paper does not permit to give the details of the calculations (it takes a forthcoming paper), we here therefore present the preliminary results of using the CUR values in the annotation workflow, which we defined to be performed in the following order:

1. Automatic text annotation with the tagger TransDict look-up,
2. Automatic exclusion of the top 10 of always syncretical multi-tags from disambiguation,
3. Automatic disambiguation of the rest of the multi-tags based on concept usage relevancy (CUR) values,
4. Manual postediting of the resulting annotations.

In assessing the conceptual annotation accuracy we used the temporal post-editing effort quantitative measure [39]. Participants in the evaluation experiment were the project members who acquired the platform knowledge and students of the South Ural State University (Russia) enrolled in a translation studies program and familiar with the computational linguistics concepts and post-editing techniques. They were given same-size portions of raw and automatically annotated texts (stage 3 output of the annotation workflow) and were asked to report on the time they spent on producing the gold annotations of the raw and automatically annotated texts. To make the evaluation less subjective, the raw and automatically annotated texts given to each participant were taken from different corpora. The reported time values were then summarized and normalized. The results showed that the post-editing time spent on the automatically annotated texts was on average 35% less than the time needed to conceptually annotate the raw text, which shows our annotation framework to be viable.

5. Conclusions

We have presented an interoperable platform for multi-grain annotation of multilingual domain corpora. The platform is a stand-alone PC application realized for Windows in the C++ programming language. The interoperability of the platform is provided by the tagset that includes conceptual information specified in the language-independent domain ontology and a universal tagging algorithm. The latter is defined to consist of two main successive procedures: ontological analysis (text-to-ontology mapping) and multi-tag disambiguation, for which both the rule-based NLP technique and/or quantitative measures can be applied. The paper covers the platform general design and its application for the conceptual annotation of the "Terrorism" domain corpora in English Russian and French. The potential of the developed interoperable platform as a research tool to define quantitative metrics for tag disambiguation is also demonstrated on the example of the conceptual-level annotation. The suggested quantitative metrics account for a) the frequency of the concept usage in unilingual corpora annotations and b) the variety of the unilingual lexical units mapped into a multilingual ontological concept. The specificity of the approach is that a) the unit of the ontological analysis is taken to be a multicomponent phrase rather than a single word and b) tag disambiguation can supported by the rule-based NLP technology through the fully functional platform tagger interpreter and/or by quantitative measures. The case study assessment of the conceptual tagging effort with the suggested an-notation workflow steps and quantitative tag disambiguation measures (without rule-based disambiguation) showed on average the 35% gain in tagging time, which proves the legitimacy of the proposed interoperable multilingual annotation frame-work. We are fully aware that more research should be done on disambiguation metrics and see it as our future work. In parallel, we will proceed with enlarging both the depth and the breadth of the multilingual ontology and the coverage of the onto-lexicons both in terms of size and the number of languages.

6. References

[1] L. Stojanović, N. Stojanovic, J. Ma. On the Conceptual Tagging: An Ontology Pruning Use Case. WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007, pp. 344–350.

[2] P. Buitelaar, S. Ramaka. Unsupervised ontology-based semantic tagging for knowledge markup. Proceedings of the Workshop on Learning in Web Search at Learning in Web Search at 22nd International Conference on Machine Learning, Bonn, Germany, 26-32, 2005.
[3] E. Charniak, M Berland. Finding Parts in Very Large Corpora. In Proceedings of the 37th Annual Meeting of the ACL, 1999, pp. 57–64P.
[4] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, G, Flake, G. Using Web Structure for Classifying and Describing Web Pages. In Proc. of the 11thWWW Conference, pp. 562–569, ACM Press, 2002
[5] L. Reeve, H. Hyoil. Survey of Semantic Annotation Platforms. In SAC '05, pp. 1634–1638, ACM Press, NY, USA, 2005, ISBN 1-58113-964-0
[6] A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff. Semantic Annotation, Indexing, and Retrieval. Elseviers Journal of Web Semantics, Vol. 2, 2005, No. 1, http://www.ontotext.com/kim/semanticannotation.html.
[7] P. A. Kogut and W. Holmes. AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages.
[8] H. Gauch. Scientific method in practice. Cambridge : Cambridge University Press. 435 P, 2003.
[9] V. Ceausu and S. Despr`es. Learning Term to Concept Mapping Through Verbs: A Case Study. Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM2007) located at the 4th International Conference on Knowledge Capture (KCap 2007), Whistler, British Columbia, Canada, October 28-31, 2007.
[10] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Roberts, A. Setzer. Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics. – Vol. 42 (5), 950–966, 2009.
[11] L. Tenenboim, B. Shapira, P. Shoval. Ontology-Based Classification of News in an Electronic Newspaper. International Book Series "Information Science and Computing", 89–97, 2008.
[12] U. Inyaem, Ch. Haruechaiyasak, Ph. Meesad, D. Tran. Ontology-Based Terrorism Event Extraction Proceedings of the 1st International Conference on Information Science and Engineering. P. 912–915, 2009
[13] A.V. Dobrov, Dobrova N. L., Soms N. L., Chugunov A.V. Semanticheskij analiz novostnyh soobshchenij po teme «Elektronnye uslugi»: opyt primeneniya metodov ontologicheskoj semantiki Trudy XVIII ob"edinennoj konferencii «Internet i sovremennoe obshchestvo» (IMS-2015). 120–125, 2015. (in Russian)
[14] Djemaa M., Candito M., Muller Ph., Vieu L. Corpus annotation within the French FrameNet: a domain-by-domain methodology. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 2016, Portorož, Slovenia 3794–380.
[15] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography 3 (4), 235–244, 1990.
[16] E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, A. Peters. Modelling Multilinguality in Ontologies. Proceedings of COLING 2008, Companion volume – Posters and Demonstrations. 67–70.
[17] M. Espinoza, A. Gómez-Pérez, E. Mena E. Enriching an Ontology with Multilingual Information. The Semantic Web: Research and Applications. ESWC Lecture Notes in Computer Science. – Springer, Berlin, Heidelberg. – Vol. 5021. 333–347, 2008.
[18] M. Chaves, M and Trojahn C. (2010). Towards a Multilingual Ontology for Ontology-driven Content Mining in Social Web Sites, 2010, – URL: https://goo.gl/sZKmS2(19.02.2021).
[19] E. A. Alatrish, D. Tošić, N. Milenkov N. Building Ontologies for Different Natural Languages. Building Computer Science and Information Systems. – Vol. 11(2). 623–64, 2014.
[20] D. W. Embley, S. W. Liddle, D. W. Lonsdale, Y. Tijerino. Multilingual Ontologies for Cross-Language Information Extraction and Semantic Search, 2019. https://pdfs.semanticscholar.org/6884/41a96b6da61295c7df39b70db2f28531370a.pdf
(last accessed 21.02.2021).
[21] S. Nirenburg, V. Raskin V. Ontological Semantics. MIT Press, Cambridge, 2004.
[22] I. Niles, A. Pease. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the 2003 International Conference on Information and Knowledge Engineering, 412–416.

[23] R. Arp, B. Smith, A.D. Spear. Building Ontologies with Basic Formal Ontology. MIT Press, Cambridge, 2010.
[24] D. Elworthy. Tagset design and inflected languages. In 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL), From Texts to Tags: Issues in Multilingual Language Analysis SIGDAT Workshop, 1–10, 1995.
[25] J. Nivre, I. Boguslavsky, L. Iomdin. Parsing the SynTagRus treebank of Russian. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 641–648, 2008.
[26] M. Gnasa, J. Woch. Architecture of a knowledge based interactive Information Retrieval System. 2002, http://konvens2002.dfki.de/cd/pdf/12P-gnasa.pdf (last accessed 19.02.2021).
[27] A. Feldman, H. Jirka, Ch. Brew. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In Procedings. LREC 2006.
[28] S. Petro, D. Das, R. McDonald. A universal part-of-speech tagset. Proceedings of the Conference on Language Resources and Evaluation (LREC 2012). Istanbul, Turkey, 2012.
[29] T. Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10), Valletta, Malta, 2010.
[30] G. Orosz, A. Nov´ak, G. Pr´osz´eky. Lessons Learned from Tagging Clinical Hungarian. IJCLA, vol. 5, no. 1, 129–145, 2014.
[31] H. F. Hsieh. Three Approaches to Qualitative Content Analysis / H.-F. Hsieh, S.E. Shannon // Qualitative Health Research. Vol. 15 (9). – P. 1277–1288, 2005.
[32] Carvalho V.A., Almeida J.P.A., Fonseca C.M., Guizzardia G.(2017). Multi-level ontology-based conceptual modeling. Data & Knowledge Engineering.Volume 109, 3-24, 2017.
[33] Ménard P.A,, Barrière C. PACTE : a collaborative platform for textual annotation –URL: https://www.aclweb.org/anthology/W17-7410.pdf (last accessed 2021/03/08).
[34] Stenetorp P., Pyysalo S., Topic G., Ohta T., Ananiadou S., Jun’ichi Tsujii J. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp.102–107. Avignon, France, 2012.
[35] Zagorul'ko M. Yu., Kononenko I. S., Sidorova E. A. Sistema semanticheskoj razmetki korpusa tekstov v ogranichennoj predmetnoj oblasti [System for Semantic Annotation of Domain-Specific Text Corpora]. Proceeding of the international conference Komp'yuternaya lingvistika i intellektual'nye tekhnologii, Bekasovo, May 30 – June 3, 2012. Moscow, RSUH, vol. 11(18), 674–683. 2012. (in Russian)
[36] P. Stenetorp, S. Pyysalo, G., Topic, T. Ohta, S. Ananiadou, J. Tsujii. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, April 23-27, Avignon, France, 102–107, 2012.
[37] S. Sheremetyeva. Universal Computational Formalisms and Developer Environment for Rule-Based NLP. Lecture Notes in Computer Science, vol 10761. Springer, Cham, 2018. https://link.springer.com/chapter/10.1007/978-3-319-77113-7_5DOI https://doi.org/10.1007/978-3-319-77113-7_5
[38] S. Sheremetyeva, A. Zinovyeva. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content. Communications in Computer and Information Science, 859. Springer, Cham, pp. 368–379, 2018.
[39] A. Zaretskaya, M. Vela, P. Corpas, M. Seghiri. Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors. New Voices in Translation Studies. 15,pp. 63-91, 2016.