Proceedings of the 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation (ISA-16), pages 102–109
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Towards Creating Interoperable Resources for Conceptual Annotation of Multilingual Domain Corpora

Svetlana Sheremetyeva
South Ural State University
76, Lenin pr. 454080, Chelyabinsk, Russia
lanaconsult@mail.dk; sheremetevaso@susu.ru

Abstract

In this paper we focus on creation of interoperable annotation resources that make up a significant proportion of an on-going project on the development of conceptually annotated multilingual corpora for the domain of terrorist attacks in three languages (English, French and Russian) that can be used for comparative linguistic research, intelligent content and trend analysis, summarization, machine translation, etc. Conceptual annotation is understood as a type of task-oriented domain-specific semantic annotation. The annotation process in our project relies on ontological analysis. The paper details on the issues of the development of both static and dynamic resources such as a universal conceptual annotation scheme, multilingual domain ontology and multipurpose annotation platform with flexible settings, which can be used for the automation of the conceptual resource acquisition and of the annotation process, as well as for the documentation of the annotated corpora specificities. The resources constructed in the course of the research are also to be used for developing concept disambiguation metrics by means of qualitative and quantitative analysis of the golden portion of the conceptually annotated multilingual corpora and of the annotation platform linguistic knowledge.

Keywords: annotation resources, conceptual domain annotation, interoperability, multilingualism, terrorism

1. Introduction

The importance of linguistic annotations and, especially, semantic annotations over raw textual data is widely acknowledged as critical in developing language technologies, such as intelligent content and trend analysis, classification, machine learning, summarization, machine translation, etc. (Mair, 2005; Pustejovsky, 2012). However, and this is also widely recognized, annotated corpora are quite sparse and their availability is often problematic due to no or restricted access, differences in volume and principles of construction, non-standardized and/or unsuitable annotations for specific language technology tasks. There are good reasons for this, - annotating a comprehensive corpus with semantic representations is a hard, costly and time-consuming task. In spite of quite a number of attempts to facilitate the problem by developing reusable annotations, including semantic annotation formats, such as, for example, XML, SGML, etc., and the introduction of increasingly convivial and hardware-independent application software, it is difficult to find a system that matches exactly end-user requirements. For quality semantic annotation, the portable annotation software packages, as the main dynamic annotation resource should contain a significant amount of linguistic knowledge, acquisition of which so far is highly problematic. If, however, genericity is considered as applied to a family of applications, i.e., applications sharing tasks and domains, one can probably suggest particular approaches to solve the problem, even cross linguistically. In this paper we attempt just that.

Our ultimate goal is to develop a methodology for developing annotation resources and resources themselves for the conceptual annotation of multilingual domain corpora, which are interoperable across languages and targeted to the automation of the annotation process primarily, but not exclusively, for such tasks as intelligent content analysis, machine learning, and classification. In our project, conceptual annotation is understood as a type of domain-specific task-oriented semantic annotation as opposed to the annotation with high level semantic properties, such as animacy, being human, person, etc.

We demonstrate our approach on the domain of e-news on terrorist attacks in three languages, English, French and Russian. Our motivation to focus on the domain on terrorist attacks is that counterterrorist activity requires, among others, operative analysis of unstructured e-information and the availability of means to speed up the creation of annotated corpora in this particular domain is of high importance. We here focus on the development of both static and dynamic annotation resources such as a universal conceptual annotation scheme, multilingual domain ontology and annotation platform with flexible settings. The platform is multipurpose; it can be used for the automation of the conceptual resource acquisition and of the annotation process itself, as well as for the documentation of the annotated corpora specificities. The resources constructed in the course of the research are also to be used for developing concept disambiguation metrics by means of qualitative and quantitative analysis of the golden portion of the conceptually annotated multilingual corpora and of the annotation platform linguistic knowledge.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work. Section 3 defines the research tasks and introduces our data set. In Section 4 we suggest a methodology of building interoperable domain-specific conceptual annotation resources and describe the pool of static and dynamic resources built in the course of the current phase of the research. Section 5 describes the "first-machine-then-human" workflow of the conceptual annotation procedure. We conclude with the research overview and future work.

2. Related Work

Today the area of language annotation research witnesses the tendency towards semantization and, in particular, domain semantization (in our research, domain conceptualization), as the most realistic way to solve language technology tasks. The current trend is to use domain ontologies as conceptual annotation instruments, which, in turn, boosts the research in the field of ontology development. Ontologies are most often created for the annotation of unilingual (most often, English) domain corpora oriented to particular tasks. For example, to name just a few, the ontology described in (Roberts A. et. al., 2009) is created for the analysis of English medical records. (Tenenboim L et al., 2008) present the domain ontology for personalized filtration of English eNews. (Mannes and Golbeck, 2005; Najgebauer et al., 2008; Inyaem et al., 2009) devote their efforts to building ontologies for forecasting terror attacks and extraction of terrorist events from eNews. There is much less research on the ontology-based annotation in other languages, among which (as most closely related to our research) are (Dobrov et al., 2015) who suggest ways to semantically annotate a Russian domain corpus, and (Djemaa et al., 2016) focusing on a French corpus, correspondingly,

As ontology development is a very tough and time-consuming task, there are attempts to save effort in constructing ontological resources by making them multilingual. Multilingualism in ontologies is generally understood in two major senses: 1) as the adaptation (or understandability) of the ontology labels for the users-native speakers of different national languages and 2) as the capability of one ontology to be applied to processing texts in different languages regardless of the language used for wording concept labels. These understandings of ontology multilingualism directly depend on the interpretation of ontology either as a language-dependent or language-independent resource.

Language-dependent ontologies are thesaurus-like structures whose elements are defined by the properties of a specific language. A well-known example of such resource, often called ontological, is the famous WordNet thesaurus (Miller et al., 1990). The research on providing ontological multilingualism here goes in the direction of localization of the labels of ontology concepts, rather than modification of the ontological conceptualization. The localization procedure can go in different ways. For example, (Montiel-Ponsoda et al., 2008) propose the association of word senses in different languages to ontology concepts through a special linguistic model, while (Espinoza et al., 2008) suggest translating ontology labels into the user's language. One more localization technique is to manually annotate ontological concepts with labels in different languages (Chaves and Trojahn, 2010). (Alatrish et al., 2014) direct their efforts to the development of universal tools that could be used for semi-automatic procedure of building separate ontologies tuned to different languages. (Embley et al., 2019) suggest methodologies on how to relate unilingual ontologies by mapping both the data and the meta-data of these ontologies. The use of language-dependent ontologies for interoperable semantic (conceptual) annotation of multilingual corpora does not seem quite doable.

Language-independent ontologies, like e.g., Mikrokosmos (Nirenburg and Raskin, 2004), SUMO (Niles et al., 2003) and BFO (Arp et al., 2015), allow multilingualism in the second sense (the applicability to processing texts in multiple languages) per definition, provided that each lexical unit (one- or multi-component) in the vocabulary of a particular language is mapped (according to special rules) into such ontology concept. This is the basic feature that makes language-independent ontologies applicable to semantic (conceptual, including) annotations that can be interoperable across languages. Given the expense of manual work, unavoidable in semantic (conceptual) annotation a lot of effort in using language-independent ontologies as annotation instruments is currently devoted to the creation of different tools to increase annotators' productivity. As a rule, so far, such annotation tools are user interfaces for mapping lexical units into ontological concepts and/or postediting the results of the automated annotation (Zagorul'ko et al., 2012; Stenetorp et al., 2012).

3. Approach and Data

3.1 Task Definition

Creation of interoperable resources for annotation makes should be closely associated with the annotation procedure that in our research is defined by the intersection of the following criteria: (i) data-driven methodology directed from analysis to representation, (ii) domain orientation, (iv) interoperability across languages, (v) automation of the annotation process, (vi) reusability of resources. We argue that interoperability of content annotation across languages calls for a clear division between language-dependent lexical knowledge and language-independent conceptual knowledge that can be best represented in ontology. We consider ontological analysis as a main instrument for interoperable conceptual annotation with a tagset defined by the ontological concepts. We are fully aware that ontological analysis has a serious limitation that lies in its practical realization. The shortcomings of ontological analysis are well-known and include the difficulty of clearly specifying the boundaries of the analysis and the influence of objective human judgments. There is no universal recipe for ideal ontological analysis therefore, as a rule, in every practical project, specific approaches are developed to deal with the problems above. Our solutions are domain-constraint and data-driven. Then, to reduce manual work, a decision was made to experiment as much as possible with the "first-machine, then human" set-up of annotation work and to postpone the actual annotation process till later stages of the research and to first focus on the creation of the resources for annotation, which, following the classification given in (Witt et al., 2009) are divided into static and dynamic. In our research static resources include a conceptual annotation knowledge that consists of multilingual comparable domain corpora on terrorist acts in three languages (English, French and Russian), a universal conceptual annotation scheme, a multilingual domain ontology, domain-related unilingual lexicons and lexical-ontological mappings. The dynamic resources are tools to automate the creation of both static resources, and the annotation procedure.

The road map for this research is as follows. First, the data set for the study was acquired and conceptualized resulting into lists of conceptually classified lexical items, and then the upper-level ontology and representation formalism were decided on followed by the development of a seed multilingual ontology for the terrorist attack domain. The seed ontology was further refined and populated with the text template technique. In parallel with the research on the content (knowledge) side of the project, a toolkit to automate the work on all its stages was being developed.

3.2 Data Set

First of all, the advantage was taken of the previously built domain resources created for our earlier CAT project that include a 400 000 word Russian terrorist domain corpus of 2016-2017 e-news acquired in the Internet and a Russian-English lexicon of multicomponent lexical units built over the corpus. The lexicon includes initial corpus-based Russian vocabulary translated into English by professional translators. These data were used, first of all, to acquire knowledge for the built-in house Internet crawlers to automatically collect new portions of Russian, as well as English and French domain corpora. The crawler knowledge was decided to consist of key phrases rather than single words as the use of key phrases has the immediate effect in improving precision in keyword related tasks (Lefever et al., 2009). Then, a general opinion that content resides in noun phrases (Witschel, 2005), made us vote in favor of keywords/phrases as grammatically well-formed noun phrases. The key noun phrases were automatically extracted from the "old" 400 000 word Russian corpus by means of the tool described in (Sheremetyeva, 2012) that we trained for the Russian terrorist domain. The top 30% of the extracted Russian key noun phrases and their translations into English and French were used as the knowledge for the crawlers, by means of which the second part of the raw data, - multilingual terrorist act corpora of 100,000 words published on the Internet in 2018-2019 in the three languages were automatically acquired. For feasibility reasons, we excluded news on terrorist military activities and focused on the news on terrorist attack committed by individuals or terrorist groups in different countries.

4. Building Resources for Annotation

4.1 Static resources

In this section we describe the process of acquisition of static resources for conceptual annotation at the pre-annotation stage. The results of the acquisition were used as the knowledge base for the NLP annotation platform (see section 4.3) and were further augmented in the course of the whole research period.

4.1.1 Data set analysis

The first step in building resources for interoperable conceptual annotation consisted in classifying the multilingual corpora lexis into domain-relevant conceptual classes (or categories). It included decisions on i) the units of conceptual classification, which we took to be both single words and multi-word phrases of different POS classes, and ii) the list of categories/concepts. The set up for this work included an initial intuitively prescribed universal list of conceptual classes with definitions, unilingual (English, French and Russian) corpora-based frequency lists of multicomponent noun phrases as most closely content-related textual units (note, not only key phrases), raw corpora for context check, if needed, and conceptualization guidelines that were the same throughout the languages. The lists of noun phrases for conceptualization were constructed in two takes. First, the set of noun phrases up to four components long[1] were automatically extracted from the English, French, and Russian corpora with the lexical extractor (Sheremetyeva, cf.) after it was trained for the terrorist domain in all the three languages. Then, every unilingual corpus was searched for longer noun phrases with the regular "find" functionality using the seed set of automatically extracted 4-component phrases. The domain-relevant units where then manually classified into conceptual classes (starting with the prescribed set) and following the guidelines. Special attention was paid to the selection of concept labels that were worded in English and made as descriptive as possible. Throughout the whole research period, weekly discussions were held by the project participants to provide for inter-conceptualization consistency and brush up. This stage resulted in the specification of the seed set of domain concepts. Other types of phrases were then extracted and classified in the same way followed by further brush up and extension of the cross-language conceptual class set. In general, the concept set was elaborated to specify a 3-level tree-like structure of concept organization with 97 fine-grained conceptual categories, assigned to 20 top-level domain categories. Table 1 shows a fragment of the top level domain concept list with definitions; Table 2 lists the second level grained concepts for the top domain concepts COUNTER-TERRORISM and CONSEQUENCES, and Table 3 presents fragments of unilingual lexica lists assigned to the conceptual class "AGENT – TERRORIST".

AGENT – TERRORIST: Executor of a terrorist act
ASSUMPTION: Assumption on who could commit a terr. act
CAUSE: What caused a terrorist attack
CLAIM RESPOSSIBILITY: terr. act responsibility claims
CONSEQUENCES: Aftermath of the terrorist attack
COUNTER-TERRORISM: People and measures against terr.
GOAL OF ATTACK: Demands of terrorists
LOCATION: Place where a terrorist act was committed
MEANS OF ATTACK: Items used for a terrorist act
NATION: person citizenship or country related to terrorism
OBJECT OF ATTACK: Who or what was hit in terr. act
TIME: Date and time when the terrorist attack happened
TYPE OF ATTACK: shooting, explosion, stubbing, arson
SOURCE : Sources of attack reports: newspapers, TV, etc.

Table 1: A fragment of the domain conceptual class list.

COUNTER-TERRORISM
COUNTER-TERRORISM AGENT : People fighting terr.
COUNTER-TERRORISM MEASURES: counter-terr. action
CONSEQUENCES
PUBLIC LOSS: killed, wounded, hostage, no damage
DESTRUCTION: objects damaged or destructed
TERRORISTS' LOSS: suicided, killed, wounded, detained
TERRORISTS' GAIN: terrorists' demands answered
PUBLIC REACTION: manifestation of support
RECONSTRUCTION: restoration of destroyed objects

Table 2: Second level concepts for the top concepts COUNTER-TERRORISM and CONSEQUENCES.

[1] Constraint to four component extraction units is explained by the limitations of the extractor.

Language	Most frequent domain lexica of the class
English	terrorist, militant, fighter, gunman, suicide bomber, jihadi, female suicide bomber, female terrorist, lone-wolf terrorist, ISIS terrorist
French	terroriste, kamikaze, combattant, femme kamikaze, djihadiste. loup solitaire, terroriste de l'EI, combattant terroriste, femme terroriste
Russian	террорист, боевик, смертник, террорист-смертник, террористка-смертница, игиловец, террористка, джихадист, террорист-одиночка

Table 3: Fragments of the most frequent unilingual lexical units put into the "AGENT-TERRORIST" class.

Like any work on semantics based on human judgment, the concept specification process, in spite of all the domain constraints and guidelines, was not free from different levels of detalization, overlaps in interpretation and even contradictions. In such cases, reasonably strict decisions were taken by the project leader.

4.1.2 Ontology

In our project, we follow three basic methodological assumptions on ontology definition. The first is that ontology is a reusable language-independent resource; the second is that "domain-specific knowledge is not isolated from general world knowledge" (Moreno & Pérez, 2011, p. 233) and we, therefore, link our ontological resource to the upper-level Mikrokosmos ontology (Nirenburg & Raskin, cf.) to reuse the knowledge that is already there. We also follow the initial Mikrokosmos division of the reality into OBJECTS, EVENTS, and PROPERTIES, and use its formalism. We keep concept labels worded in English, the scopes of which, like in Mikrokosmos, are only specified by their definitions. Our third assumption is that interoperable domain ontological knowledge can be extracted from multilingual comparable domain corpora using mixed (top-down/bottom-up) acquisition techniques (Francesconi et al., 2010).

The set of domain concepts defined at the lexical analysis stage formed the seed e-news terrorist ontology, whose pool of concepts was further augmented and refined by using the text-template technique. For example, such RELATION concepts as IS-A and INSTANCE-OF can be acquired (though not exclusively) using the following English/French/Russian parallel text templates:

"A / is / are / and other/ such as/ B" (English)

"A est /somme/ comme / et autres / B" (French)

"A / это / и другие / такие как / B" (Russian),

where B is a lexeme that can signal of a more general concept; A is a lexeme of a more specific class.

The top domain concept MEANS OF ATTACK can be further split by means of such corsslingual templates as

"attack /*with/ using/involving*/ C" (English)

"attaque /*avec/au moyen de*/ C" (French)

"атака /*с использованием/ с применением*/ C" (Russian),

where C stands for lexemes of a weapon type concept.

The resulted ontology currently consists of 112 OBJECT and EVENT concepts and 27 PROPERTY concepts, see details in (Sheremetyeva & Zinovyeva, 2018).

4.1.3 Lexical-Ontological Mapping

Our main methodology for the interoperable conceptual annotation is ontological analysis. In practice, ontological analysis consists in mapping corpora lexical units into ontological concepts that, in our case, calls for creating unilingual lexicons, in which every domain-related unit is explicitly linked to an ontological concept. The boundaries of such mappings were specified by the domain data analysis and where allowed to be one-to-many, many-to-one or many-to-many. This had to follow human judgement, though strictly regulated by the mapping guidelines. For example, the French named entity "Charlie Hebdo" is mapped into the concepts OBJECT OF ATTACK (its office was targeted by terrorists in 2015) and SOURCE (it is a weekly newspaper that published info on terrorist attacks). Among lexical items mapped to several concepts there are, for example, the English word "police officer" and its French and Russian equivalents "policier" and "полицейский", correspondingly. Namely, following their use in the corresponding unilingual corpus, these lexical items are mapped into the 4 concepts of the multilingual ontology (the order of examples below are English, French and Russian):

COUNTER-TERRORISM: After the explosions, the authorities deployed *police officers*. / Après les explosions, les autorités ont déployé des *policiers*. / После взрывов власти выставили *полицейских*.

CONSEQUENCES: A *police officer* was killed. / Un *policier* est tué. / *Полицейский* был убит.

SOURCE: According to a *police officer*, the man shouted "Allahu akbar". / Selon des *policiers*, l'homme aurait crié « Allah akbar ». / По словам *полицейских* мужчина кричал «Аллах акбар»

AGENT-TERRORIST: Russia's ambassador is assassinated in Ankara by a *police officer*. / L'ambassadeur de Russie est assassiné à Ankara par un *policier*. / Российский посол убит в Анкаре *полицейским*.

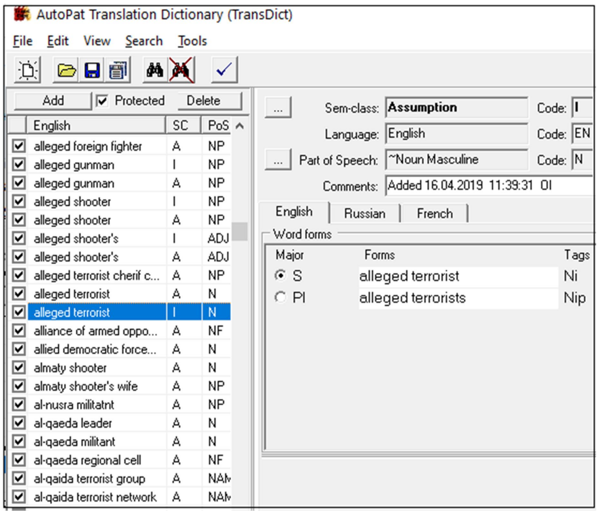
We also introduced a convention that is not very obvious and generally accepted. It concerns the ontology mapping of multicomponent lexical units, in which individual components bear domain-related conceptual meanings that translate different aspects of content and do not contradict one another. For example, in the English phrase "airport shooting suspect", the word "shooting" conveys the information on the type of attack, the word "airport" points to the location where the attack took place, while the word "suspect" has two conceptual meanings "assumption" and "performer of the terrorist attack". All these content components are sincretically united in the phrase. Therefore, the convention is to map this multi-component lexeme into 4 concepts, - AGENT-TERRORIST, ASSUMPTION, TYPE OF ATTACK and LOCATION. Similarly, the phrase "Algerian terrorist" is mapped into the AGENT-TERRORIST and NATION concepts. Multiple ontological-lexical mappings will obviously lead to assigning multiple concept tags to textual units in the annotation procedure. However, as seen from the examples above, in our approach to annotation, it might or might not signal of lexical unit conceptual ambiguity. The situation forecast the need to make decisions on when multiple conceptual tags have to be disambiguated and when it should not be done to preserve as much domain-related content as possible. This issue is a matter of further investigation.

The domain conceptualization described in this section resulted in the acquisition of the pre-annotation static knowledge including the multilingual terrorist act domain ontology and ontology-mapped corpora-based unilingual lexicons of English, French and Russian. This knowledge was used to create the first version of the multilingual annotation platform described in Section 4.2.

4.2 Annotation Platform

A tool, which we call annotation platform, is the main part of all dynamic resources we used in our work. We approached its design with several considerations in mind. First of all, the annotation platform should automate the process of conceptual annotation and mark-up every unilingual corpus with the universal set of concept tags defined by the multilingual domain ontology. It is also desirable for the platform to contain knowledge that could help conceptual disambiguation. Further, it should be possible to configure the platform settings to different languages and language-dependent types of linguistic information. The annotation platform should allow for the knowledge administration and, therefore, be provided with the acquisition interface.

To save the development effort we reused, though sufficiently updated two software modules from our earlier (different type) project (Sheremetyeva, 2013) that meet most of the expectations on the annotation platform. The first module is the program shell of the multilingual TransDict e-lexicon and the second is the tagger to which TransDict is pipelined. TransDict is built over a powerful set of linguistic features that have a tree-like structure. It is realized as a number of cross-referenced monolingual lexicons. Every monolingual lexicon consists of a set of entries with semantic, syntactic and morphological zones of flexible settings. The TransDict entry is meant for one meaning (semantic class) of a lexeme in a given language. The morphological zone can contain the morphological information, such as part-of-speech, number, gender, etc., and word paradigms of a lexical unit up to 10 components long explicitly listed in the entry. The latter makes recognition of text wordforms straightforward. Depending on the configuration of linguistic information, every wordform in the lexicon entry is automatically assigned a supertag that codes semantic and morphological information, such as part-of speech and typed morphological features that are language-dependent. TransDict, what is important for our project, has an advanced knowledge administration user interface, built-in search module with flexible search masks and a lot of other effort-saving functionalities, like automatic generation of entry structures and entry-fillers. The TransDict shell allows increasing the number of languages as necessary and can be configured to any type of knowledge. The adaptation of TransDict for the conceptual annotation task (see Figure 1) was as follows. We configured the program to three languages, - English, French and Russian. Semantic classes were set to the ontology concepts and some other classes like "Other" "Numerals", "Definiteness", etc., for mapping the lexemes of not specifically domain-related meaning. For feasibility reasons, so far, only upper-level ontology concepts were coded in TransDict. The morphological zones of the entries within each conceptual class were filled up with the explicitly listed morphological paradigms of the lexemes mapped to the ontology at the pre-annotation static resource acquisition stage (see Section 4.1.). If a lexical unit was mapped to several conceptual classes, several entries for this unit were created, each linked to a particular concept. Figure 1 shows a fragment of the TransDict main acquisition interface with the word list filtered by the mask "English" & "mapped into the TERRORIST-AGENT (tag A) concept" & "also to any other concept". The duplication of the lexical units shown in the left column of the interface displays multiple mappings. For example, the two-component lexeme "alleged terrorist" is listed twice as it is mapped into the TERRORIST-AGENT concept (tag A) and into the ASSUMPTION concept (tag I).



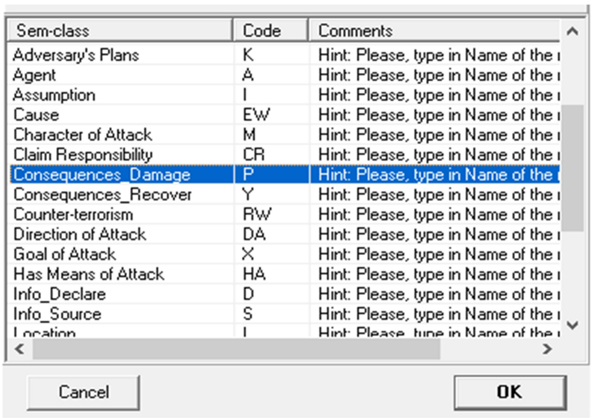
	English	SC	PoS
☑	alleged foreign fighter	A	NP
☑	alleged gunman	I	NP
☑	alleged gunman	A	NP
☑	alleged shooter	I	NP
☑	alleged shooter	A	NP
☑	alleged shooter's	I	ADJ
☑	alleged shooter's	A	ADJ
☑	alleged terrorist cherif c...	A	NP
☑	alleged terrorist	A	N
☑	alleged terrorist	I	N
☑	alliance of armed oppo...	A	NF
☑	allied democratic force...	A	N
☑	almaty shooter	A	N
☑	almaty shooter's wife	A	NP
☑	al-nusra militatnt	A	NP
☑	al-qaeda leader	A	N
☑	al-qaeda militant	A	N
☑	al-qaeda regional cell	A	NF
☑	al-qaida terrorist group	A	NAM
☑	al-qaida terrorist network	A	NAM

Sem-class: Assumption Code: I
Language: English Code: EN
Part of Speech: ~Noun Masculine Code: N
Comments: Added 16.04.2019 11:39:31 OI

Major	Forms	Tags
● S	alleged terrorist	Ni
○ Pl	alleged terrorists	Nip

Figure 1: A fragment of the main TransDict interface.

The screenshot displays the "alleged terrorist" lexeme entry mapped into the ASSUMPTION concept. The morphological zone is filled with the lexeme wordforms that are automatically assigned supertags Ni and Nip, where N stands for "noun", "I" for the concept ASSUMPTION and "p" for plural. Supertags are positional, a concept code is the second in order; this coding format is inherited from the parent TransDict application. To allow the acquirers working independently at their own pace, TransDict is programmed in two variants, as MASTER with a full set of functionalities and as the so-called SLAVE – an empty program shell configured exactly as the corresponding version of MASTER but of a limited capability, namely, the user cannot change the dictionary settings (sets of languages, conceptual classes, entry structures and tags). SlAVEs filled by the acquirers with new portions of lexical conceptual knowledge are merged into MASTER on a regular basis. TransDict entries can be created for a single lexeme or for whole lists in batch mode. Figure 2 shows the window for ontological mapping when a lexeme is to be added to the TransDict knowledge. The window pops-up following a click on the "Add" button in the interface.

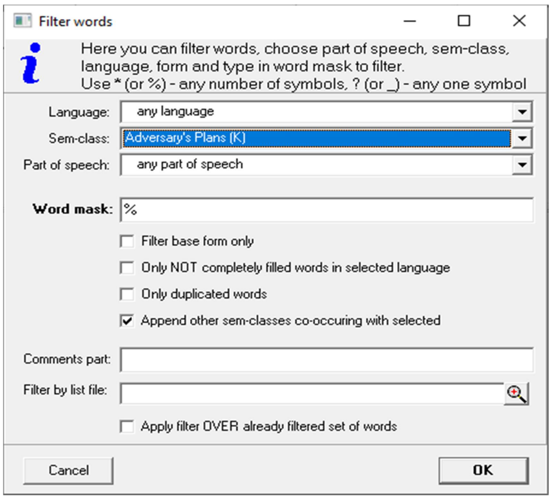


Sem-class	Code	Comments
Adversary's Plans	K	Hint: Please, type in Name of the
Agent	A	Hint: Please, type in Name of the
Assumption	I	Hint: Please, type in Name of the
Cause	EW	Hint: Please, type in Name of the
Character of Attack	M	Hint: Please, type in Name of the
Claim Responsibility	CR	Hint: Please, type in Name of the
Consequences_Damage	P	Hint: Please, type in Name of the
Consequences_Recover	Y	Hint: Please, type in Name of the
Counter-terrorism	RW	Hint: Please, type in Name of the
Direction of Attack	DA	Hint: Please, type in Name of the
Goal of Attack	X	Hint: Please, type in Name of the
Has Means of Attack	HA	Hint: Please, type in Name of the
Info_Declare	D	Hint: Please, type in Name of the
Info_Source	S	Hint: Please, type in Name of the
Location	L	Hint: Please, type in Name of the

Figure 2: TransDict pop-up window for lexical-ontological mapping and assigning concept tags.

The selection of a conceptual class calls for another pop-up window for part-of-speech specification, after which an entry with typed morphological and syntactic zones appears that could be filled, if and as necessary. Fillers of the TransDict morphological zone fields supply knowledge to the tagger for conceptual annotation.

As said above, original TransDict shell was substantially updated for the annotation knowledge management and now includes quite a number of new effort-saving acquisition and analysis functionalities, substantially augmented search/filtering possibilities, export/ import functions, etc. The new TransDict search module with a lot of possible search masks is shown in Figure 3. The main update here is filtering according to the concept class parameters (combined or not with other mask parameters). One can filter lexemes of one conceptual class, lexemes of one class that are also mapped to any other concepts, and lexemes assigned to a fixed set of conceptual classes. This function shows knowledge lacuna to be filled. Filtration on the concept parameters can be done in two modes: based on lexeme main forms only or based on the whole paradigm of lexeme wordforms listed in the TransDict morphological zone. This obviously gives different results, comparing which one can find morphological hints for concept disambiguation in an automatically annotated text.



Here you can filter words, choose part of speech, sem-class, language, form and type in word mask to filter.
Use * (or %) - any number of symbols, ? (or _) - any one symbol

Language: any language
Sem-class: Adversary's Plans (K)
Part of speech: any part of speech

Word mask: %

□ Filter base form only
□ Only NOT completely filled words in selected language
□ Only duplicated words
☑ Append other sem-classes co-occuring with selected

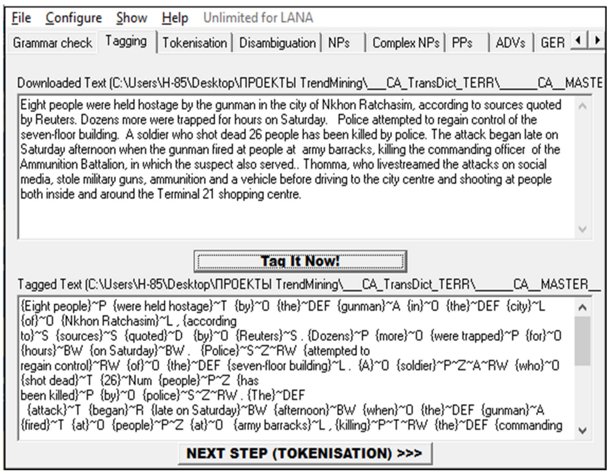
Comments part:
Filter by list file:

□ Apply filter OVER already filtered set of words

Figure 3: TransDict pop-up window for lexical-ontological analysis.

In general, all types of filtering including the concept class masks give a lot of information on the domain annotation statistics that can be used e.g., for forecasting the conceptual ambiguity rate in a particular language and for developing automatic disambiguation metrics.

The second module of the annotation platform is the tagger pipelined to TransDict. The tagger has a control interface and compilers which, if necessary, can be used for the acquisition of disambiguation rules and syntactic analysis rules. The tagger can be set to coarse-grain or fine-grain corpus mark-up. The coarse-grain mark-up outputs annotation with concept tags only, which can be enough for certain text-mining and content/knowledge extraction tasks. The fine-grain mark-up assigns a full range of linguistic features coded in the TransDict supertags that can be useful for disambiguation purposes. A screenshot of the control interface of the annotation platform tagger with the results of coarse-grain automatic conceptual tagging is shown in Figure 4 (see the concepts tags in Figure 2). Some lexemes shown in the tagger interface screenshot have multiple tags that signals of possible conceptual ambiguity. This version of the tagger does yet support concept disambiguation and, in general, the problem of automated conceptual disambiguation is out of the scope of this paper. We can only say at this stage that both statistical and, if necessary, linguistic information will be used for this purpose. This, among others, motivated the main change in the current tagging module as compared to the parent application, - two level fine-grained and coarse-grained annotation.



Downloaded Text (C:\Users\H-85\Desktop\ПРОЕКТЫ TrendMining__CA_TransDict_TERR_____CA__MASTE

Eight people were held hostage by the gunman in the city of Nkhon Ratchasim, according to sources quoted by Reuters. Dozens more were trapped for hours on Saturday. Police attempted to regain control of the seven-floor building. A soldier who shot dead 26 people has been killed by police. The attack began late on Saturday afternoon when the gunman fired at people at army barracks, killing the commanding officer of the Ammunition Battalion, in which the suspect also served.. Thomma, who livestreamed the attacks on social media, stole military guns, ammunition and a vehicle before driving to the city centre and shooting at people both inside and around the Terminal 21 shopping centre.

Tagged Text (C:\Users\H-85\Desktop\ПРОЕКТЫ TrendMining__CA_TransDict_TERR_____CA__MASTER__

```
{Eight people}~P  {were held hostage}~T  {by}~O  {the}~DEF  {gunman}~A  {in}~O  {the}~DEF  {city}~L
{of}~O  {Nkhon Ratchasim}~L , {according
to}~S  {sources}~S  {quoted}~D  {by}~O  {Reuters}~S . {Dozens}~P  {more}~O  {were trapped}~P  {for}~O
{hours}~BW  {on Saturday}~BW .  {Police}~S~Z~RW  {attempted to
regain control}~RW  {of}~O  {the}~DEF  {seven-floor building}~L .  {A}~O  {soldier}~P~Z~A~RW  {who}~O
{shot dead}~T  {26}~Num  {people}~P~Z  {has
been killed}~P  {by}~O  {police}~S~Z~RW . {The}~DEF
 {attack}~T  {began}~R  {late on Saturday}~BW  {afternoon}~BW  {when}~O  {the}~DEF  {gunman}~A
{fired}~T  {at}~O  {people}~P~Z  {at}~O  {army barracks}~L , {killing}~P~T~RW  {the}~DEF  {commanding
```

Figure 4: The tagger interface showing the results automatic coarse-grained conceptual annotation.

The annotation platform is currently implemented as a PC application and includes three pipelined modules, TransDict MASTER, TransDict SLAVE and Tagger that can also be used as stand-alone tools.

5. Annotation Procedure

In our approach, the process of conceptual annotation as the implementation of ontological analysis is the process of mapping text strings (in our case grammatical phrases of different types) into the domain multilingual ontology. The annotation procedure is identical for each unilingual corpus. It is “first-machine, then human” and is incremental in nature. We tested the approach, given the volume of multilingual effort and expectations about reasonable annotator tasks, on relatively small portions of unilingual corpora of 20,000 words each. However, the process and the results of such annotation, which in the long run was potedited into golden, gave us a lot of experience and leads on how to treat conceptual annotation problems.

During the beginning annotation phases covered in this paper, the types of conceptual categories included in the annotation were constrained to 21 top-level domain concepts and the concept "OTHER", to which domain-neutral lexemes are mapped. The annotators, who had already been trained in conceptualizing during the lexical analysis stage, were given a code-book with the sets of concepts associated with definitions and tags. The annotation process itself was done in several takes in an iterative manner. First, a weakly portion of the raw text meant to be gold-annotated was automatically tagged by our annotation platform described in Section 4.2 and then passed for postediting to the annotators. Conceptual ambiguity, if any, was resolved manually. In case a domain-relevant lexeme was left untagged or tagged incorrectly, it was supplied with correct linguistic information into the acquirer's personal TransDict SLAVE program to be further merged in TransDict MASTER (see Section 4.2) and the platform knowledge was thus updated, after which the annotation platform was used to automatically annotate the next portion of the corpus leading to a new knowledge update, etc. The knowledge was updated on a regular basis and the accuracy of the automatic annotation increased with very iteration. The accuracy was so far evaluated based on the annotators' reports on the amount of time spent on postediting and on the number of new lexical items to be merged into TransDict after every annotation iteration. Evidently, one cannot hope for a 100% correct automatic annotation without some risk of reducing annotation quality and, hence, human judgements cannot be avoided. However, our experiment shows that automation as used in the current research significantly augments and supports the annotation process.

The annotation procedure resulted in three golden conceptually annotated comparable English, French and Russian corpora of the e-news on terrorist acts and a substantial augmentation of the annotation platform knowledge. The TransDict lexicon currently consists of three unilingual lexicons of the English, French and Russian languages, that amount to around 43000 cross-referenced lexical entries acquired both at the pre-annotation stage, and in the course of annotation.

6. Conclusion

In this paper, we suggested a methodology of creating static and dynamic resources for interoperable conceptual annotation of domain corpora and presented actual annotation resources built along the suggested methodology for the multilingual (English, French and Russian) domain corpora of e-news on terrorist attacks. The resources include a universal conceptual annotation scheme, multilingual domain ontology, annotation platform with flexible settings and comparable golden conceptually annotated corpora in the three languages. This research is one of the major parts of an annotation project, which is significantly different from those that concentrate on morphological, syntactic or general types of semantic annotation. The emphasis of the presented work is on: i) a domain-specific level of annotation; ii) the assignment of well-defined interoperable conceptual representations based on multilingual domain ontology; and iii)"first-machine-then-human" approach to the annotation process.

Qualitative and quantitative investigation of the annotation resources we have constructed open quite a number of research opportunities for, e.g., theoretical aspects of social and comparative linguistics, as well as for research and development in Natural Language Processing technologies including multilingual Information Extraction, Generation, Question Answering, etc., and Machine Translation. The conceptual annotation knowledge can directly be used for developing machine learning techniques. In particular, the resource analysis findings can be used for developing concept disambiguation metrics, which, on top of increasing the volume of the annotation resources and annotated corpora, we see as our future work.

Bibliographical References

Alatrish E.A., Tošić D., Milenkov N. (2014). Building Ontologies for Different Natural Languages. Building Computer Science and Information Systems. – Vol. 11(2). pp. 623–644.

Arp, R., Smith, B., Spear, A.D. (2010). Building Ontologies with Basic Formal Ontology. MIT Press, Cambridge.

Chaves, M and Trojahn C. (2010). Towards a Multilingual Ontology for Ontology-driven Content Mining in Social Web Sites – URL: https://goo.gl/sZKmS2(09.11.2019).

Djemaa M., Candito M., Muller Ph., Vieu L. (2016). 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 pp. 3794–3801.

Dobrov, A.V., Dobrova N.,L., Soms N., L., Chugunov A.V. (2015). Semanticheskij analiz novostnyh soobshchenij po teme «Elektronnye uslugi»: opyt primeneniya metodov ontologicheskoj semantiki Trudy XVIII ob"edinennoj konferencii «Internet i sovremennoe obshchestvo» (IMS-2015). pp. 120–125. (in Russian).

Embley D. W., Liddle S. W., Lonsdale D. W., Tijerino Y. (2019). Multilingual Ontologies for Cross-Language Information Extraction and Semantic Search. – URL: https://pdfs.semanticscholar.org/6884/41a96b6da61295c7df39b70db2f28531370a.pdf ((09.11.2019)

Espinoza, M., Gómez-Pérez A., Mena E. (2008). Enriching an Ontology with Multilingual Information. The Semantic Web: Research and Applications. ESWC Lecture Notes in Computer Science. – Springer, Berlin, Heidelberg. – Vol. 5021. pp. 333–347.

Francesconi E., Montemagni S., Peters W., Tiscornia D. (2010). Integrating a Bottom-Up and Top-Down Methodology for Building Semantic Resources for the Multilingual Legal Domain Semantic Processing of Legal Texts . LNAE. – Vol. 6036, pp. 95–121.

Inyaem U, Haruechaiyasak Ch., Meesad Ph,,Tran D. (2009). Ontology-Based Terrorism Event Extraction

Proceedings of the 1st International Conference on Information Science and Engineering.. – P. 912–915.

Lefever E., Macken L.. Hoste V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus'. In Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece. pp. 496–504.

Mannes, A., Golbeck J. (2005). Building a Terrorism Ontology/ Proceedings of the ISWC Workshop on Ontology Patterns for the Semantic Web 36. URL: https://pdfs.semanticscholar.org/9bcb/90e48677e39da7b84939e8c8da2b2a63cde7.pdf (25.09.2019).

Mair, C. (2005). The corpus-based study of language change in progress: The extra value of tagged corpora. The AAACL/ICAME Conference, Ann Arbor, 2005.

Montiel-Ponsoda E., Aguado de Cea G., Gómez-Pérez A., Peters A. (2008). Modelling Multilinguality in Ontologies. Proceedings of COLING 2008, Companion volume – Posters and Demonstrations. pp. 67–70.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J. (1990). Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography 3 (4), pp. 235–244.

Moreno A., Pérez Ch. (2011). From Text to Ontology Extraction and Representation of Conceptual Information. Actes de quatrièmes rencontres «Terminologie et Intelligence Artifi-cielle», pp.233-242.

Najgebauer A., Antkiewicz R., Chmielewski M., Kasprzyk R., (2008). Prediction of Terrorist Threat on the basis of Semantic Association acquisition and Complex Network Evolution. The Journal of Telecommunications and Information Technology. Vol. 2. pp. 14–20.

Niles I. & Pease A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), pp. 412–416.

Nirenburg S. & Raskin V. (2004). Ontological Semantics. MIT Press, Cambridge

Pustejovsky J. (2012). Natural Language Annotation for Machine Learning. O'Reilly Media; 1 edition. 342 P.

Roberts A., Gaizauskas R., Hepple M., Demetriou G., Guo Y., Roberts A., Setzer A. (2009). Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics. – Vol. 42 (5), pp. 950–966.

Sheremetyeva S. (2012). Automatic Extraction of Linguistic Resources in Multiple Languages. Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, Poland, pp. 44–52.

Sheremetyeva S. & Zinovyeva A. (2018). On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content. Communications in Computer and Information Science, 859. Springer, Cham, pp. 368–379.

Sheremetyeva S. (2013). On Integrating Hybrid and Rule-Based Components For Patent MT with Several Levels of Output. Proceedings of "The Fifth Workshop on Patent Translation in conjunction of the fourteenth Machine Translation Summit 2013", Nice, France, September 2-6.

Stenetorp P., Pyysalo S., Topic G., Ohta T., Ananiadou S., Jun'ichiTsujii J. (2012). 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. Avignon, France, April 23 – 27. 2012. pp. 102–107.

Tenenboim L., Shapira B, Shoval P. (2008). Ontology-Based Classification of News in an Electronic Newspaper. International Book Series "Information Science and Computing", pp. 89–97.

Witschel H. F. (2005). Terminology extraction and automatic indexing - comparison and qualitative evaluation of methods. Terminology and Knowledge Engineering (TKE) http://wortschatz.unileipzig.de/~fwitschel/papers/TKEIndexing.pdf

Witt, A., Heid, U., Sasaki, F., Gilles Sérasset (2009). Multilingual language resources and interoperability. Lang Resources & Evaluation 43, 1–14 (2009). https://doi.org/10.1007/s10579-009-9088-x

Zagorul'ko, M. YU., Kononenko I. S., Sidorova E. A. (2012). Sistema semanticheskoj razmetki korpusa tekstov v ogranichennoj predmetnoj oblasti. Proceeding of the international conference Komp'yuternaya lingvistika i intellektual'nye tekhnologii, pp. 674–683. (in Russian).