On Resolving Conceptual Ambiguity in an English Terrorism E-news Corpus

Anastasia Zinoveva [0000-0002-7658-7376]

South Ural State University, 76 Pr. Lenina, Chelyabinsk 454080, Russia
zinovevaaiu@bk.ru

Abstract. Conceptual ambiguity in a restricted domain is a crucial, yet under-investigated issue in ontological analysis as it complicates the process of extracting relevant information from unstructured texts. This paper aims to reveal the sources of conceptual ambiguity in an English terrorism e-news corpus and explore ways to resolve it. In our study, empirical corpus-based methods are employed to determine sources of conceptual ambiguity in the corpus and suitable disambiguation methods. Our findings reveal four sources of conceptual ambiguity, namely, part-of-speech homography, lexical ambiguity, the plurality of conceptual meanings, and the extralinguistic context. We analyze three quantitative corpus-based methods applied to different types of conceptual ambiguity and outline the prospects of future research in this area.

Keywords: Ontological analysis, Conceptual annotation, Conceptual ambiguity, English e-news corpus, Terrorism.

Introduction

Research on ontological analysis is crucial for the development of natural language processing technologies witnessing now a trend towards semantization of textual metadata based on ontological analysis. Ontological analysis, which can be defined as the process of eliciting content knowledge of entities involved in a certain domain [5], essentially consists in, firstly, annotating lexical items in a text with ontology concept tags and, secondly, formalizing and interpreting the results of such annotation in accordance with a specific task. The first part of this process can be called (broadly) semantic annotation [6, 15], or, more precisely, concept labeling [14] or conceptual annotation, the term we use in this project to represent the notion of domain-oriented ontology-based semantic annotation.

Semantic annotation is a technique used to enrich content with various semantic information. Depending on the interpretation of this notion, it can be approached in three different ways: the first interpretation defines it as the process of assigning particular senses to polysemantic words usually based on a dictionary or an ontology [3]; according to the second interpretation, it means attributing certain universal semantic features to words based on a certain lexical classification [10]; in the third interpretation, it is viewed as mapping semantic relations between words in a text [9].

Conceptual annotation, in turn, can be viewed as a special case of semantic annotation based on ontology concepts and, perhaps, domain-oriented. Although similar to the first interpretation of semantic annotation in terms of mapping lexical items into an ontology, it is not entirely the same as its results are relevant in a specific domain; e.g., *people* in terrorism e-news regularly represent consequences of an attack (as they can be killed, injured, etc.), although this meaning is not inherent in the word.

Conceptual annotation can be done manually [6, 15], semi-automatically [13] or automatically [14]. Although manual conceptual annotation of text corpora ensures high-quality results if done properly, it requires considerable time and rigorous training of annotators to avoid inconsistencies, hence the need to facilitate rapid conceptual annotation by automating it to a certain extent. Insofar as processing of large corpora is concerned, semi-automatic annotation also has its limitations in terms of timing and training of annotators. Attempts at complete automation, in turn, cause various challenges, conceptual ambiguity in particular [14], even in a restricted domain [12].

We define **conceptual ambiguity** as a problem that emerges when a lexical item is assigned two or more (possibly, mutually exclusive) concept tags in the process of automatic conceptual annotation. Conceptual ambiguity seems relatively easy for an annotator to resolve; however, for a computer, it is irresoluble unless specific disambiguation instructions are given. To the best of our knowledge, little research has addressed this issue (see [14] where a machine learning algorithm was used to resolve conceptual ambiguity in web queries); nonetheless, the task of conceptual ambiguity resolution is close to Word Sense Disambiguation (WSD), which is "the ability to identify the meaning of words in context in a computational manner" [7], with allowance that, instead of the meaning of a word, the ontological concept is the subject of identification, which makes WSD methods, both knowledge-based and corpus-based, potentially applicable for conceptual ambiguity resolution as well. Other methods of corpus linguistics can also be applied for domain-oriented conceptual ambiguity resolution, namely the Edmundsonian methods for key word identification [4].

To this end, this paper addresses the problem of conceptual ambiguity in automatically tagged English terrorism e-news aiming to determine the sources of conceptual ambiguity in the corpus and explore corpus-based methods for disambiguation.

The rest of the paper is divided into three parts: Section 2 describes the resources used for ontological analysis, as well as the annotation scheme and procedure; in Section 3, the results of the experiment are presented and discussed; in Section 4, a conclusion is made and research prospects are outlined.

1 Resources & Experiment Procedure

1.1 Resources for ontological analysis

The experiment described in this article involved several resources:

– a language-independent domain ontology designed to process terrorism-related e-news in three languages: English, French, and Russian (see the development details in [11, 12]);

– an English lexicon linked to the ontology;
– a software prototype for conceptual annotation;
– a raw English terrorism e-news corpus (11,296 words).

Our domain ontology is represented in the MikroKosmos formalism, with its division of the reality into OBJECTS, EVENTS, and PROPERTIES (further divided into RELATIONS and ATTRIBUTES) [8], and contains 112 OBJECT and EVENT concepts and 27 PROPERTY concepts. Table 1 shows some of the top ontology concepts with their definitions. It should be noted that in order to ensure interoperability between the domain ontology and the MikroKosmos, concept labels are worded in English, though the content of a particular concept is determined by its definition, rather than by its label. For example, the scope of the concept WEAPON is not limited to weapons only as it can be seen from its definition 'weapons or weapon-like objects used to commit a terror attack, also functional weapon parts'. Thus, such lexical items as *nail bomb*, *explosive*, *truck*, *bullet*, etc. will all be mapped into WEAPON in the ontology.

Table 1. Some of the terrorism ontology concepts with definitions

Concept	Definition
AGENT	The perpetrator or organizer of an attack or an organization behind it.
WEAPON	Weapon or weapon-like objects used to commit terror attacks, also functional weapon parts.
TERROR ATTACK	An attack committed by a terrorist or a group of terrorists to intimidate population and achieve ideological or political goals.
LOCATION	The place where a terror attack was committed.

Although the ontology is multilingual and is linked to three lexicons, in this study, we apply it to the ontological analysis of an English corpus in search of language-specific indicators for conceptual ambiguity resolution. The English lexicon is composed of lexical items of up to 10 components; some of them are shown in Table 2.

Table 2. English lexical items mapped into ontology concepts

Concept	Lexical items
AGENT	accomplice of a suicide bomber, adversary, former soldier, infamous militant, jihadist gunman, knife-wielding man
WEAPON	armored car bomb, assault rifle, bomb-laden vehicle, combustible liquid, homemade mortar, incendiary mixture, lorry, vest
TERROR ATTACK	attempted hijacking, deadly shooting rampage, explosion, hostage taking, intimidation act, knife attack, mass shooting
LOCATION	downtown, fast food restaurant, Quetta hospital, railway station

The software prototype for conceptual annotation is based on lexical and ontological knowledge and designed to annotate texts with ontology concept tags. The raw English corpus was tagged automatically with this tool and post-edited manually to resolve conceptual ambiguity and obtain its "golden" version. The automatically tagged and "golden" corpora were then compared to reveal the sources of conceptual ambiguity.

1.2 Tagging schema and procedure

For feasibility, we selected 22 top ontology concepts and coded them under the tags: A = AGENT, BW = TIME, C = WEAPON, CR = CLAIM RESPONSIBILITY, D = DECLARE, DA = DIRECTION, E = OTHER TERRORIST ACTIVITIES, EW = CAUSE, HA = HAVE WEAPON, I = ASSUMPTION, K = ADVERSARY'S PLANS, L = LOCATION, M = SCALE OF ATTACK, N = NATION, OW = OTHER, P = CONSEQUENCES, RW = COUNTER-TERRORISM, S = SOURCE, T = TERROR ATTACK, UW = TERRORIST ORGANIZATION, X = GOAL OF ATTACK, Z = OBJECT OF ATTACK. We also used a number of tags for some lexical items irrelevant for the domain, at least in some contexts: different kinds of predicates (B, R, U), noun phrases (PO), adjectives, adverbs, names and abbreviations (O), numbers (Num), unknown items (UNK), and determiners (DEF). That was done to avoid linking a word to a concept in a non-terrorist context.

Normally, one lexical item should be assigned one tag at a time, with some notable exceptions that can be viewed as manifestations of conceptual syncretism. We define **conceptual syncretism** as a possibility for a lexical item to be mapped simultaneously into two or more concepts in the "golden" corpus. For example, in the sentence *At least 15 people were killed in an explosion that hit the rebel-held city of al-Bab in northern Syria*, tagging Syria with N and L is justified as it is both NATION and LOCATION, and both of the concepts are equally meaningful in the context[1]. The noun *suspect* is another good example of conceptual syncretism as it is tagged both A (AGENT) and I (ASSUMPTION), because a suspect, per definition, is a person suspected of a crime, i.e. a possible agent. However, if *suspect* is a verb, the tag A will be superfluous making this a case of part-of-speech homography (see Section 3.3 for details).

These examples show that conceptual syncretism is different from conceptual ambiguity in the fact that it does not need to be resolved; moreover, conceptually syncretic lexical items represent potential subconcepts or concepts modified by properties. For instance, in the above example, the combination L-N means that NATION can be linked to some other concept (TERROR ATTACK, which is clear from the context) by means of the relation LOCATION-OF. The combinations S-N (e.g., *Turkish media*) or Z-N (e.g., *Iranians*) mean that the SOURCE of the message or the OBJECT of the attack belong to certain nations/ethnic communities. The combination A-I, in turn, indicates that someone is assumed to be the AGENT, but it is not confirmed. Hence, ASSUMPTION can be viewed as an attribute able to modify other concepts.

However, only two concepts (N and I) are able to freely form tag combinations as manifestation of conceptual syncretism according to our current rules. Some other concepts form tag combinations with restrictions: e.g., RW can be a part of a tag combination unless its constituents are mutually exclusive (*RW-K, *RW-T), K can be paired with T (e.g., *to prepare an attack*) or OW (e.g., *to plan to recruit terrorists*), Z with L (e.g., *mosque*), UW with A (if a terrorist organization is not only mentioned in the text but also acts as the agent of a particular attack), etc.

[1] The order of tags is not meaningful and is only determined by algorithms of our annotation tool.

2 Results & Discussion

2.1 General corpus analysis results

The study found 193 unique tags in the automatically tagged corpus, 163 of them tag combinations (or multitags). After post-editing, 45 unique tags remained, 15 of them multitags that represent conceptual syncretism. Furthermore, we calculated relative frequencies of tag occurrences in both corpora and determined that the ratio of all (conceptual and non-conceptual) multitags to all tags in the automatically tagged corpus was 24 % and the ratio of conceptual multitags to all conceptual tags was 43 %, while in the "golden" corpus, the respective rates were significantly lower, 4 and 9 %.

Fig. 1 and 2 show the respective distributions of multitags in the automatically tagged corpus and its "golden" counterpart. Although some of the tags (e.g., L-N, P-RW, A-UW) appear in both figures, most of the tags in Fig. 1 require disambiguation. The data obtained suggest that conceptual ambiguity is rather frequent in the domain texts; moreover, it is diverse, and various methods might be needed to resolve it.

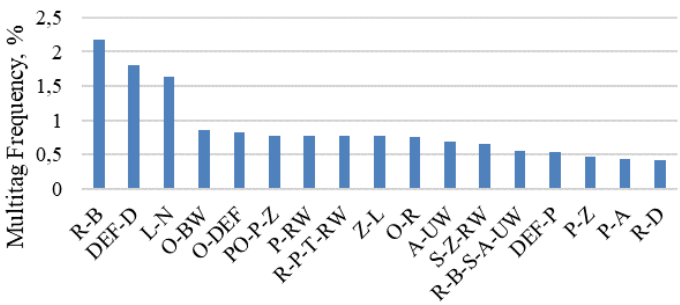

Fig. 1. Distribution of 17 most frequent multitags in the automatically tagged corpus; 100 % is the total number of tag occurrences in the corpus

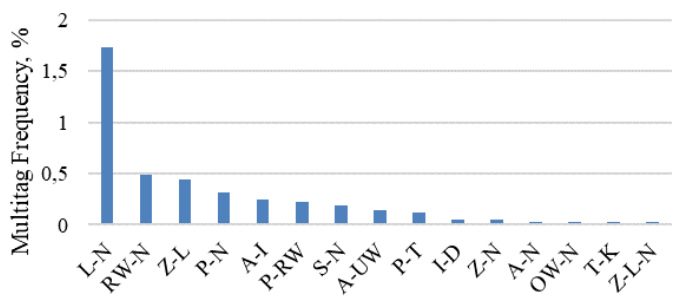

Fig. 2. Distribution of multitags in the "golden" corpus; 100 % is the total number of tag occurrences in the corpus

2.2 Sources of conceptual ambiguity

To identify the sources of conceptual ambiguity, we performed a comparative study of the automatically tagged corpus and the "golden" corpus. As a result, four types of conceptual ambiguity depending on the source have been revealed.

Part-of speech (POS) homography. This type of ambiguity arises when lexical items are identical at least in one form but they belong to different parts of speech. The examples are quite numerous in the English corpus among both domain-relevant and domain-irrelevant one-component lexical items: {act}$^{\sim T\sim R}$, which should be disambiguated as T (TERROR ATTACK) if *act* is a noun or as R if *act* is a non-conceptual verb; {bomb}$^{\sim T\sim C}$, which should be disambiguated as T (TERROR ATTACK) if *bomb* is a verb and as C (WEAPON) if *bomb* is a noun; {suspect}$^{\sim A\sim I}$, which should be left as is if *suspect* is a noun or disambiguated as I (ASSUMPTION) if *suspect* is a verb. Other examples are {is}$^{\sim R\sim B\sim S\sim A\sim UW}$, {may}$^{\sim BW\sim I}$, {sat}$^{\sim R\sim BW}$, {said}$^{\sim DEF\sim D}$, {report}$^{\sim S\sim D}$, {us}$^{\sim O\sim S\sim Z\sim L\sim RW\sim N}$. It should also be mentioned that lexical items in forms which do not coincide with each other (e.g., *has been reported*, *had bombed*, *suspected*) are ascribed a single tag due to the functionality of our annotation tool.

Lexical ambiguity. It is a possibility for a lexical item to have two or more interpretations in the context, which can be caused either by homonymy or by polysemy. For instance, the word *release* can have one of the two lexical meanings in the corpus: 'to make free' (about hostages or terrorists) or 'to make public' (about statements). Although choosing one meaning in the context is easy for annotators (cf. *Hostages were released* vs. *A statement was released*), it is not so for the annotation tool, hence the multitag RW-D (COUNTERTERRORISM / DECLARE). Some other examples are[2]:

be directed (DIRECTION OF ATTACK / OTHER TERRORIST ACTIVITIES)
1) 'to be aimed at an object' (about terror attacks)
2) 'to be guided by advice, helpful information' (about terrorists)

body (CONSEQUENCES / COUNTERTERRORISM)
1) 'a corpse'
2) 'a collective group'

station (LOCATION / SOURCE)
1) 'a bus or train station'
2) 'a radio or television channel'

underground cell (TERRORIST ORGANIZATION / COUNTERTERRORISM)
1) 'a small group acting as a unit within a larger terrorist organization'
2) 'a small room in prison located underground'

Khorasan (LOCATION / TERRORIST ORGANIZATION)
1) 'a region in the Middle East'
2) 'a branch of the Islamic State located in the Khorasan region'
3) 'an alleged group of senior al-Qaeda members operating in Syria'

[2] The definitions are taken from Dictionary.com and Wikipedia.org and adjusted to the terrorism domain if needed.

Plurality of conceptual meanings. It is manifested in one-to-many mappings between lexical items, which are identical in form and dictionary meaning, and ontology concepts, with only one of the latter relevant in a specific context. This can be illustrated by the word *police*: the word appears in three distinct types of sentences in the corpus:

1. **Police** (= COUNTERTERRORISM) apprehended the suspect.
2. The attacks targeted **police** (= OBJECT) and the military.
3. **Police** (= SOURCE): 5 dead, 8 wounded in airport shooting.

Thus, it is automatically annotated with a combination of tags S-Z-RW, only one of which should be preferred in each case — the decision that is quite easy for an annotator to make, but difficult for a computer. Ambiguities of this type are the most frequent in the corpus among both one- and multicomponent lexical items (see Table 3).

Table 3. Lexical items with plural conceptual meanings.

Lexical item/Tag	A	P	RW	S	T	Z
authorities			●	●		
detonated			●		●	
government forces		●	●			●
government official		●		●		●
fighter	●	●	●			
foreign tourist		●				●
killing		●	●		●	
military		●	●	●		●
soldier	●	●	●			●

Extralinguistic context. Ambiguities of this source are the hardest, is possible, to resolve even by an annotator. They arise from extralinguistic differences such as attitudes of different parties towards a certain issue. For example, the phrase *a Dogon group* can be tagged both A (AGENT) and RW (COUNTERTERRORISM) due to the fact that the Dogon militia is considered responsible for several attacks against the Peuhl community in Mali, which, in turn, is accused by Dogons of sympathizing with Islamist militants. In this perspective, Dogons' actions can be viewed as counterterrorism.

Each case of ambiguity induced by the extralinguistic context requires a thorough examination, after which only one point of view must be adopted, be it that of the global community or some other source. However, corpus data may not be enough to resolve this kind of ambiguity and commonsense knowledge should be employed in disambiguation. Considering the infrequency of this type of ambiguity in the corpus and the complexity of obtaining commonsense knowledge of this level, we currently find it extraneous to resolve ambiguity induced by the extralinguistic context.

It is important to mention that a multitag can emerge from several sources at the same time: e.g., the multitag for the word *accused* is R-P-A-I-D, wherein the plurality of conceptual meanings and POS homography are intertwined. If *accused* is a substantivized adjective, only the A-I tags are relevant; otherwise, if *accused* is

a verb, one of the R-P-I-D tags should be preferred. Cases of mixed-source conceptual ambiguity are not infrequent in the corpus, and they should be resolved progressively.

2.3 Disambiguation methods

In this section, we investigate three quantitative corpus-based methods for conceptual ambiguity resolution in the terrorism domain:

- a tag-ranking-based method;
- a co-occurrence-based method;
- a positional method.

Some other potentially useful methods are also paid attention to. In this paper, knowledge-based methods are not addressed as they require additional examination.

Tag-ranking-based method. This method is loosely based on the one described in [10], where it was proposed to renumber lexical meanings of certain lexemes based on corpus data as opposed to dictionaries and establish a hierarchy of meanings to resolve semantic ambiguity in the Russian National Corpus. In our turn, we propose to build tag rankings for frequent lexical items for which tag rankings can be calculated based on corpus data.

In our corpus, 10 lexical items meet the high frequency requirement: *army*, *children*, *civilians*, *control*, *incident*, *is*, *military*, *people*, *police*, and *security personnel*. For four of them, the data are not enough to build complete rankings, e.g., the word *military* which is automatically assigned tags P-S-Z-RW appears only with two tags in the "golden" corpus — RW and Z, which means that the military do not act as SOURCE in the test corpus and CONSEQUENCES related to them are not specified. Another example is the word *is* which is automatically tagged R-B-S-A-UW, but has either R or UW in the "golden" corpus, with R-tagged items accounting for 95 %. It clearly is a case of homography (*is* as a third-person present singular form of *to be* and *IS* as an acronym for *the Islamic State*) caused by our annotation tool not distinguishing between uppercase and lowercase letters. Even though the data are incomplete, given the high relative frequency of R-tagged items and a higher rate of other acronyms (*ISIS/ISIL*), we assume that referring to the Islamic State as *IS* is uncommon in our corpus, hence the multitag R-B-S-A-UW is unlikely to be disambiguated as S, A, or UW. Tag rankings for the other six items are: *army* — RW, S; *civilians* — P, PO; *control* — E, RW; *incident* — T, PO; *police* — S, RW, Z; *security personnel* — P, RW. While in some cases the predominance of a tag is clear (*incident* is TERROR ATTACK in 88 % cases, *civilians* represent CONSEQUENCES in 86 % cases), most of them are borderline.

Apparently, this method has noticeable drawbacks. Firstly, a considerable number of lexical items in the corpus are not frequent enough to build complete tag rankings for them. Secondly, frequencies of several tags assigned to one lexical item can be equal or close to equal, and thus none of them can be preferred over the other. Therefore, this method can be used either for a very limited number of lexical items, for which one of the tags clearly prevail (so that a "primary" tag can be identified), or in

combination with some other method. The reliability of the method, alone and in combination with other methods, is to be tested in further research.

Co-occurrence-based method. To test this method, we have built concordances for each tag in the "golden" corpus using freeware corpus analysis toolkit AntConc [1] and applied them to resolve conceptual ambiguity of the multitag S-Z-RW in the sentences:

1. **{Police}**$^{\sim S\sim Z\sim RW}$ {apprehended}$^{\sim P\sim RW}$ {the}$^{\sim UNK}$ {suspect}$^{\sim A\sim I}$.
2. {The}$^{\sim UNK}$ {attacks}$^{\sim T}$ {targeted}$^{\sim DA}$ **{police}**$^{\sim S\sim Z\sim RW}$ {and}$^{\sim O}$ {the}$^{\sim UNK}$ {military}$^{\sim P\sim S\sim Z\sim RW}$.
3. **{Police}**$^{\sim S\sim Z\sim RW}$: {5}$^{\sim Num}$ {dead}$^{\sim P}$, {8}$^{\sim Num}$ {wounded}$^{\sim P}$ {in}$^{\sim O}$ {airport shooting}$^{\sim T\sim L}$.

To resolve the S-Z-RW ambiguity in these three cases, we examined the narrow context: the right one in Cases 1 and 3 and the left one in Case 2. In Case 1, S-Z-RW is followed by P-RW (the double tag here is caused by conceptual syncretism, hence no disambiguation required). The concordance shows that P-RW can be preceded by A, L, DEF, I, RW, O, and P. Since no other tag of the listed is a part of S-Z-RW, it should be disambiguated as RW. In Case 2, S-Z-RW is preceded by DA; meanwhile, according to the concordance, DA is either preceded by Z or followed by it (with rare inclusions of UNK or DEF) in all cases, hence, here S-Z-RW should be disambiguated as Z. Finally, in Case 3, S-Z-RW is followed by Num and P. In the concordance, this sequence appears only once preceded by S, while it does not appear at all with the other two tags, which gives some indications, but cannot be completely relied on.

Although this method might not produce accurate results in all cases, we believe that tag co-occurrence can be used as one of the measures to calculate the probability of conceptual ambiguity resolution.

Positional method. This method is based on certain considerations from text stylistics and its idea is close to that of the classical location method for key word identification proposed by Edmundson [4]. News articles can be structured in various manners, but one of the most frequent and effective patterns is an inverted pyramid [2], where the information is presented in descending order, with the most important points written at the beginning of an article and the least important ones mentioned at the end. For this reason, we propose a hypothesis that sentences or entire sections of an article with an inverted pyramid structure that are closer to the top may have a higher rate of terrorism-related tags than distant ones. In this perspective, lexical items annotated with terrorism-related concept tags can be viewed as key words.

Following this assumption, we divided each of the articles in the "golden" corpus in two ways: 1) into sentences; 2) into five identifiable sections: Headline, Lead, Main Story, Background, and Reactions[3]; then, we formed testing positional subcorpora

[3] The headline and the main story which describes (shortly or in detail) a terror attack or some other terrorism-related event are mandatory for any news article, while three other sections are optional. The lead paragraph is placed right after the headline and used to summarize main ideas of an article. The background section gives additional information related to the

containing all 1st, 2nd, …, 19th sentences in Case 1 and all headlines, leads, etc. in Case 2 and calculated tag frequencies for each subcorpus.

The results for Case 1 are shown in Fig. 3. One can observe a decrease in concept tag relative frequency from 62 to 28 % in sentences 1–9, but then the curve rises again and remains at around 40 % with an upsurge to 47 % in sentence 18. The increase and the upsurge can be explained by two factors: firstly, the articles in question are of different length and from different sources, hence the discrepancies in structure; secondly, several articles contain information of other terror attacks at the end, which is related to, yet not crucial for the main story, hence the higher rate of concept tags. The results for Case 2 are similar (see Fig. 4), with a decrease from 62.3 % concept tags in Headlines to 40.2 % in Backgrounds.

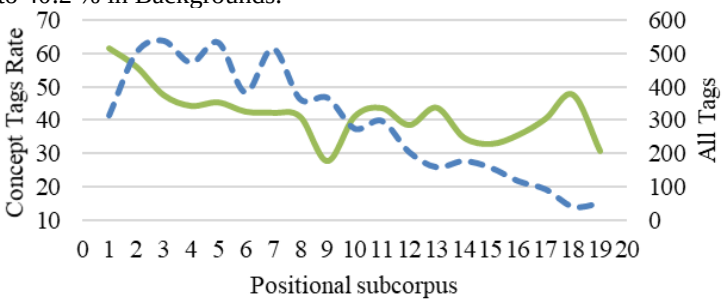

Fig. 3. Distribution of concept tags (green curve, left axis) depending on sentence position in an article (bottom axis) as compared to the number of all tags in the sentence subcorpus (dashed blue curve, right axis)

With that in mind, we can assume that the multitag PO-P-Z assigned to the word *people* may be disambiguated as PO if the sentence where it occurs is closer to the end of an article, e.g., *[There are] {**people**}*$^{\sim PO}$ *who have hatred for Islam and Islam is about peace* (Sentence 13 or Background). Meanwhile, if the sentence is closer to the beginning, the multitag can be disambiguated as either P or Z, as in *Car bomb kills 11 {**people**}*$^{\sim P\sim Z}$ *(?) in Mogadishu* (Sentence 1 or Headline). Neither P, nor Z can be preferred here so far due to the lack of specific positional data for these concepts, and some other methods should be used to disambiguate them.

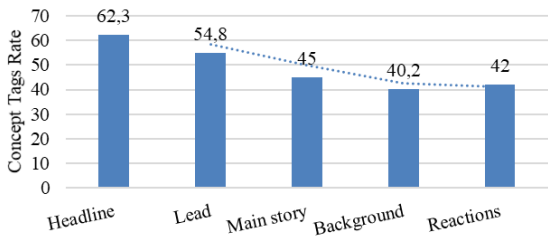

Fig. 4. Distribution of concept tags in sections of an article

attack or its perpetrator, such as the context of the attack or the terrorist's social background, while the reactions section provides various opinions regarding the attack.

It should be noted that the positional method based on sectional data has a serious drawback that makes it hard to apply in automatic ontological analysis: precisely, it might be difficult to divide the article into sections other than Headline, Lead, and Body (which comprises Main Story, Background, and Reactions) both automatically and manually, because they can be hard to distinguish; moreover, pieces of Background and Reactions can alternate throughout the article or be omitted altogether. On the contrary, sentence position data are easier to obtain and thus they seem to be of use as an auxiliary measure to resolve conceptual ambiguity.

Other methods. Methods that could potentially be used to resolve conceptual ambiguity are not limited to those already named; in particular, POS-homography-induced ambiguity can be resolved after disambiguation on the morphological level, and polysemy-induced ambiguity, after disambiguation on the lexical level. Indeed, if *act* is already determined to be a noun, it becomes clear that the right tag is T, not R; if *said* is determined to be a verb, rather than an adjective, the correct tag is D, not DEF. Furthermore, if we have already disambiguated the sense of the word *release* ('to make free' or 'to make public'), it is easy to decide whether it should be tagged P-RW or D. There is also a hypothesis to be checked that some concepts are more frequently represented in certain morphological forms (a so-called morpho-conceptual correlation), e.g., it seems logical enough that CONSEQUENCES are manifested rarely, if ever, in verbs in the future tenses. However, these methods should be further investigated.

Conclusion

We have presented the results of an experimental study of an English terrorism e-news corpus and have made an attempt to contribute into the field of domain-specific concept disambiguation. The study has shown that conceptual ambiguity must be treated as a serious enough problem of automatic ontological analysis due to its frequency and diversity both in the number of possible concept tag combinations and its sources. We have identified four sources of conceptual ambiguity that are present in the corpus to different extents; they are, namely, part-of-speech homography, lexical ambiguity, the plurality of conceptual meanings, and the extralinguistic context. We have also investigated three quantitative corpus data-based methods to resolve conceptual ambiguity and we find them rather promising, possibly when applied in combination to achieve a higher accuracy. Yet, these methods are to be further studied, tested, and evaluated on larger corpora and in other languages, as their accuracy (and even applicability) can be different for English, French, and Russian.

References

1. Anthony, L.: AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University, https://www.laurenceanthony.net/software, last accessed 2020/01/11.
2. DeAngelo, T.I., Yegiyan, N.S.: Looking for Efficiency: How Online News Structure and Emotional Tone Influence Processing Time and Memory. Journalism & Mass Communication Quarterly 96 (2), 385–405 (2019). DOI: 10.1177/1077699018792272.

3. Djemaa, M., Candito, M., Muller, Ph, Vieu, L.: Corpus annotation within the French FrameNet: a domain-by-domain methodology. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 3794–3801, Portorož, Slovenia (2016).
4. Edmundson, H.P.: New Methods in Automatic Extracting. Journal of the Association for Computing Machinery 16 (2), 264–285 (1969).
5. Façanha, R.L., Cavalcanti, M.C., Campos, M.L.M.: A systematic approach to Review Legacy Schemas Based on Ontological Analysis. In: MTSR 2018: Metadata and Semantic Research, pp. 63–75. Springer, Chamyu (2019).
6. Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9, 9–10 (2008).
7. Navigli, R.: Word Sense Disambiguation: A Survey. ACM Computing Surveys 41 (2), 69 (2009). DOI:10.1145/1459352.1459355.
8. Nirenburg, S., Raskin, V.: Ontological semantics. Cambridge: MIT Press (2004).
9. Palmer, M., Gildea, P, Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Computational Linguistics 31 (1), 71–106 (2005).
10. Rakhilina, E.V., Kobritsov, B.P., Kustova, G.I., Lyashevskaya, O.N., Shemanayeva, O.J. Semantic ambiguity as an application-oriented problem: word class tagging in the RNC. In: Computational Linguistics and Intellectual Technologies. Proceedings of the International Workshop Dialogues 2006, Moscow, pp. 445–450 (2006) (in Russian).
11. Sheremetyeva, S., Zinovyeva, A.: On modelling domain ontology knowledge for processing multilingual texts of terroristic content. In: Communications in Computer and Information Science 859. Springer, Cham, pp. 368–379 (2018). DOI: 10.1007/978-3-030-02846-6_30.
12. Sheremetyeva, S., Zinoveva, A.: Ontological analysis of e-news: a case for terrorism domain. In: Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction, pp. 130–141, Ulyanovsk (2019).
13. Song, D., Chute C. G., Tao, C.: Semantator: A Semi-automatic Semantic Annotation Tool for Clinical Narratives. In: 10th International Semantic Web Conference (2011).
14. Viju, J.S.: Concept Interpretation by Semantic Knowledge Harvesting. International Journal for Research in Applied Science & Engineering Technology (IJRASET) 6 (5), 477–484 (2018). DOI:10.22214/ijraset.2018.5081.
15. Zagorulko, M.J., Kononenko, I.S., Sidorova, E.A.: System for semantic annotation of domain-specific text corpora. In: Proceedings of the International Conference “Dialogue-2012” (2012).