A multilingual wikified data set of educational material

Hendrickx, Iris; Takoulidou, Eirini; Naskos, Thanasis; Kermanidis, Katia Lida; Sosoni, Vilelmini; De Vos, Hugo; Stasimioti, Maria; Van Zaanen, Menno; Georgakopoulou, Panayota; Egg, Markus; Kordoni, Valia; Popovic, Maja; Van Den Bosch, Antal

published in
LREC 2018 - 11th International Conference on Language Resources and Evaluation
2019

document version
Publisher's PDF, also known as Version of record

A Multilingual Wikified Data Set of Educational Material

Iris Hendrickx[1], Eirini Takoulidou[3], Thanasis Naskos[3], Katia Lida Kermanidis[3], Vilelmini Sosoni[2], Hugo de Vos[1], Maria Stasimioti[2], Menno van Zaanen[4], Panayota Georgakopoulou[6], Markus Egg[5], Valia Kordoni[5], Maja Popovic[5], Antal van den Bosch[1]

[1]Centre for Language Studies, Radboud University,Nijmegen, Netherlands

[1]Department of Foreign Languages, Translation and [2]Interpreting, Ionian University, Corfu, Greece

[3]Department of Informatics, Ionian University, Corfu, Greece

[4]Department of Communication and Information Sciences, Tilburg University, The Netherlands

[5]Department of English Studies, Humboldt−Universität zu Berlin, Germany

[6]Deluxe Entertainment Services Group, Athens, Greece

i.hendrickx@let.ru.nl;rinoulit@gmail.com;anaskos@ionio.gr; vilelmini@hotmail.com; kerman@ionio.gr; stasimioti.maria@gmail.com; mvzaanen@uvt.nl; yota.georgakopoulou@bydeluxe.com; kordonie@anglistik.hu-berlin.de; maja.popovic@anglistik.hu-berlin.de; markus.egg@anglistik.hu-berlin.de;a.vandenbosch@let.ru.nl

Abstract

We present a parallel wikified data set of parallel texts in eleven language pairs from the educational domain. English sentences are lined up to sentences in eleven other languages (BG, CS, DE, EL, HR, IT, NL, PL, PT, RU, ZH) where names and noun phrases (entities) are manually annotated and linked to their respective Wikipedia pages. For every linked entity in English, the corresponding term or phrase in the target language is also marked and linked to its Wikipedia page in that language. The annotation process was performed via crowdsourcing. In this paper we present the task, annotation process, the encountered difficulties with crowdsourcing for complex annotation, and the data set in more detail. We demonstrate the usage of the data set for Wikification evaluation. This data set is valuable as it constitutes a rich resource consisting of annotated data of English text linked to translations in eleven languages including several languages such as Bulgarian and Greek for which not many LT resources are available.

Keywords: MOOCs, crowdsourcing, wikification

1. Introduction

TraMOOC (Translation for Massive Open Online Courses) is an EU-funded project that aims to improve access to educational material in MOOCs by providing dedicated machine translation (MT) solutions. The educational content, including video lecture subtitles, forum posts, and quiz questions and answers, is translated from English into eleven European and BRIC languages: Bulgarian, Chinese, Croatian, Czech, Dutch, German, Greek, Italian, Polish, Portuguese, and Russian.

Here, we focus on the collection of parallel texts in eleven language pairs that are annotated and linked with respect to the entities occurring in the text. We consider names and noun phrases that contain the topical information about a text as entities. To identify such entities, information from Wikipedia is used. Each entity in the text is linked to its corresponding Wikipedia page in the respective language.

The fine-grained information encapsulated in this data set can be used for a wide range of applications. We use the data set for an in-depth analysis and evaluation of machine translation output. In addition to word-based evaluations (e.g., BLEU), a semantic evaluation can be performed, as the links to the Wikipedia pages in the various target languages provide an additional source of information. Such implicit MT evaluation aims to judge the MT quality between source and target language without using an explicit (manual) translation step. Alternatively, the data set is suited for other multilingual tasks that focus on semantic aspects such as the cross-lingual Semantic Textual Similarity task (Agirre et al., 2016; Cer et al., 2017). Furthermore, the data set can be used as multilingual training material for the development of novel wikification tools (e.g., Tsai and Roth (2016)), or tools that automatically detect and link topics in a text to their respective Wikipedia pages.

In the remainder of this paper we discuss related work in Section 2., the creation of this data set via crowdsourcing (Section 3.), the difficulties that we encountered using crowdsourcing for such complex annotation task (Section 4.) and the outcomes of this process in Section 5.. We also briefly discuss the use case of implicit translation evaluation for which we created the data set in Section 6. and conclude in Section 7..

2. Related Work

Comparable and related types of data sets are those created for the evaluation of wikification tools (Mihalcea and Csomai, 2007). The Illinois Wikifier was evaluated on English material annotated with Wikipedia links (Ratinov et al., 2011). This evaluation set consists of Wikipedia pages and news articles. In the context of automatic word sense disambiguation, there is another related multilingual data set from Semeval-2015 task 13 (Moro and Navigli, 2015) containing links to BabelNet (Navigli and Ponzetto, 2012) and Wikipedia pages with news articles in three languages (of which only Italian matches the languages targeted in the TraMOOC project).

These data sets do not cover all the languages we are interested in and, additionally, do not target the educational domain. Therefore, a new data set needed to be created. This data set is intended as both tuning material for the developed implicit machine translation system evaluation tool and for testing the final machine translation systems. The data set also gives us insights into the coverage of

high		low	
EN	5546	CS	397
DE	2140	BG	238
NL	1919	HR	181
IT	1407	EL	141
RU	1445		
PL	1259		
PT	987		
ZH	985		

Table 1: Number of Wikipedia articles per language in K articles (stats from 08-01-2018)

Wikipedia of the eleven project languages for topics in the educational domain, and thus testing the limits of this implicit MT evaluation method. We can only measure translation success for those topics for which an equivalent Wikipedia page exists in the target language. Tables 1 shows the number of available pages for each of the languages of the TraMOOC project. The English Wikipedia is by far the largest, consisting of more than 5,5 million articles; it is likely that many detected source topics do not have a corresponding topic in the target language. Some of our target languages have many Wikipedia pages such as Dutch and German who have at least a million articles for each language, while Greek and Croatian are supported by a much smaller number of Wikipedia pages, and we can expect this method to be less effective for them.

Due to the specialized nature of this task and its multilingual aspects, we chose to use crowdsourcing to collect the data. Even though using a non-expert crowd may perhaps lead to lower quality results, crowdsourcing has the main benefit of access to people speaking the different languages.

3. Data Selection

Our aim was to collect wikification annotations for 500 to 1,000 sentences from parallel educational texts for each language pair. We used three existing parallel text resources as the basis for the sentence selection so as to cover a broad range of online courses and to cover all eleven language pairs. In particular, we use parallel texts from course material of the Coursera MOOC platform, the Iversity MOOC platform, and the QCRI Educational Domain (QED) Corpus (formerly known as QCRI AMARA Corpus) (Abdelali et al., 2014). The Iversity data consists of (i) manually translated MOOC data, and (ii) MT output of English MOOC data produced by the first MT software prototype that was developed in the first year of the TraMOOC project. Both Coursera and QED material consist of MOOC subtitles that were translated using crowdsourcing in other projects unrelated to TraMOOC. The QED corpus consists of a large collection of files and each file contains a number of subtitles from MOOC video lectures in a particular language. The aligned files (parallel corpus) share the same first part of the file name. However, not all files with the same name contain the same amount of text. Hence, a filtering process was applied to select only those files where both the English file and the target language file contain the same number of sentences.

A mix of the data collected from the three parallel corpora was used to cover the number of sentences needed for all eleven target languages. Hence, in most of the cases both QED and Coursera material is used (with the exception of Croatian where only a small set of sentences in the Coursera corpus was available). Iversity MOOC material was available for three language pairs, i.e., EN-EL, EN-IT, EN-PT; this was combined with a sample from the Coursera corpus to reach the number of sentences required.

We deliberately chose to select sentences from different resources as each resource has its own peculiarities. For example, the QED texts include partially aligned sentences or are aligned at the clause level, while the Coursera data often aligns multiple sentences to multiple sentences. The quality of the translation and of the sentence alignment also varies per language and resource. Some sentences are badly aligned and the aligned parts do not contain the same information. This is particularly problematic in the context of multilingual wikification.

The actual text snippets used in the annotation collection task were manually selected. They consist of around 20–50 consecutive sentences each. Special care was taken to make sure that the sentences in the data set did not contain repetitions (e.g., the sentence "welcome to this lecture" is commonly used in all courses) and did not contain speaker interjections in textual representation of speech (such as "mmm", [NOISE], or [MUSIC]). Also, particularly long paragraphs were not selected, so as to maintain the microtasking nature of the activity.

The selected source and target sentences were automatically tokenized using the multilingual tokenizer Ucto[1] (van Gompel et al., 2017). Ucto has language-specific rules for the tokenization of several languages including Dutch, English, German, Italian, Portuguese, and Russian. For the other languages, generic language-independent settings of Ucto were used except for the Chinese language where we applied the Stanford Word Segmenter (Chang et al., 2008).

4. Annotation via Crowdsourcing

Crowdsourcing has been used extensively for annotating corpora due to being a cheap and fast means to collecting human intelligence input, compared to requesting expert-based intervention (Wang et al., 2013). Crowdsourcing approaches vary from voluntary work and gaming to paid microtasks (Bougrine et al., 2017). Applications involve, but are not limited to, the annotation of speech corpora (Bougrine et al., 2017; Su et al., 2017), of named entities (Bontcheva et al., 2017), of domain-specific concepts (Good et al., 2014) of relation extraction (Liu et al., 2016). Researchers have shown particular interest in issues pertaining to ethical implications (Cohen et al., 2016), as well as best practices for obtaining optimal quality output (Sabou et al., 2014). Best practice guidelines are followed in the present work also, after experimentation with varying

[1]Ucto is freely available at `https://languagemachines.github.io/ucto/`.

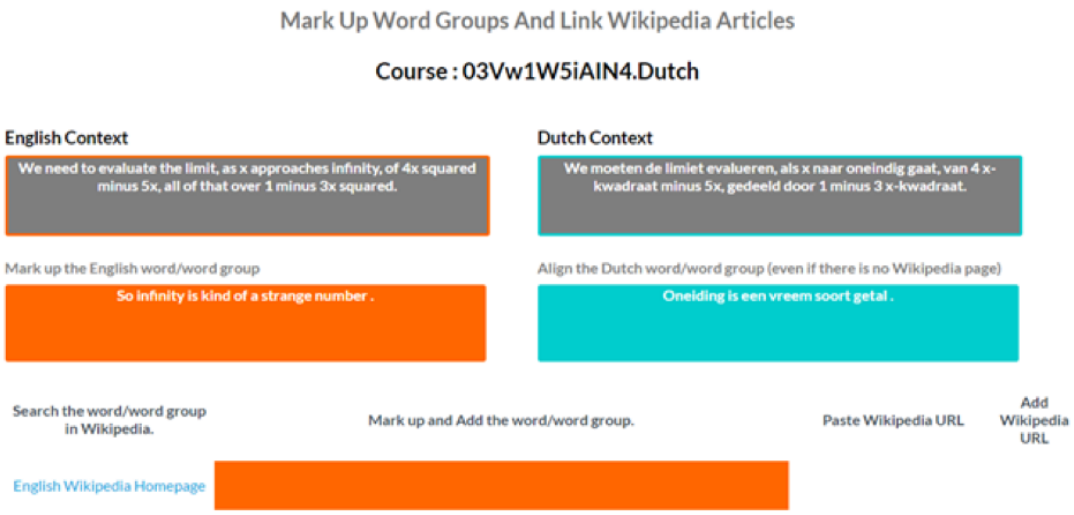

Figure 1: CrowdFlower interface for manual annotation.

parameterization schemata, and involve task decomposition into simple microtasks, the appropriate crowd choice, the appropriate crowd reward choice, data preparation, task design, task completion time, quality control, task monitoring and crowd evaluation.

Given pairs of aligned texts, the entity annotation (wikification) task consists of identifying and annotating names and noun phrases in the texts that can be linked to their corresponding Wikipedia pages with the same meaning.

To collect this information, annotators (called contributors in the context of crowdsourcing) are presented with an English sentence and its translation. Such a sentence pair is displayed along with its context (which consists of sentences of the entire paragraph in both English and the target language, as well as the related course title). Contributors are then instructed to analyze the sentences to identify names and nouns that are found in Wikipedia. If such an entity is identified, contributors mark them up in both the English and translated sentences. Annotating an item consists of highlighting the words describing the entity as well as providing a URL to the corresponding Wikipedia page in the correct language. It may be the case that an entity can be found in the English Wikipedia, but not in the Wikipedia of the target language. In this case, the phrase is still identified, but a null value is assigned for the Wikipedia link in the target language.

As this activity differs greatly from regular crowdsourcing tasks, detailed instructions with examples are provided to the contributors. Contributors were asked to mark up only those entities that had an Wikipedia page for that entity with the same meaning. In case multiple possible entities could be marked, contributors should mark the longest possible phrase.

(1) Agrippa 's Trilemma states that there are three options if we try to prove the truth .

In example 1 both [Agrippa] and [Trilemma] have their own separate Wikipedia page, but we annotate the longest word group [Agrippa's Trilemma] that points, via redirection, to the Wikipedia page entitled 'Münchhausen trilemma'. Wikipedia uses redirection links to link synonyms to the same page. In this example only two entities are to be marked, [Agrippa's Trilemma] and [truth]. The noun phrase 'options' does not express a clear concept for which a unique Wikipedia page exists and should not be marked. Such annotation decisions are sometimes difficult to make and partly subjective. For that reason each sentence is annotated by three different contributors to only keep those annotations where at least two of the three contributors agreed upon.

A trial instance ('test question') is provided to the contributor, which has gold standard annotations against which the contributor's annotation is checked. The same approach is used to validate the accuracy of the contributors' work and assure the quality of the data collected during the task. We used these test questions to monitor the trustworthiness of the contributors and flag malicious users as untrusted. These test questions were manually translated and annotated by professional translators. Around 30 test questions per language pair were prepared.

For the collection of the annotations, the CrowdFlower (CF) platform[2] was used. This choice was driven by several practical concerns, such as payment options, configurability, quality control mechanisms as well as the size of the contributor crowd. Especially the configurability of the CrowdFlower platform had a major impact on the choice of platform, as the wikification annotation task is more complex than most crowdsourcing annotation tasks.

We show a picture of the CF interface in Figure 1. The English sentence and corresponding sentence in the target language are shown next to each other in the CF interface and some preceding sentences are shown in the gray box above each sentence. Samples of consecutive sentences were presented to contributors in order to help them pick the correct Wikipedia pages for topics that have multiple meanings and, thus, multiple candidate Wikipedia links. When sentences are presented in context, this context aids the disambiguation process. Furthermore, by using several consecutive sentences, the contributors could work more efficiently

[2]CrowdFlower: `https://www.crowdflower.com`.

when encountering a topic that was already mentioned in the previous sentence.

The example sentence shown in Figure 1 was taken from the QED corpus. The short English sentence to be annotated, 'So infinity is kind of a strange number', is more easily comprehensible as part of a mathematical context when reading the preceding sentence 'We need to evaluate the limit, as x approaches infinity, of 4x squared minus 5x, all of that over 1 minus 3x squared.'. Both the English term 'number' and its Dutch translation 'getal' are ambiguous terms that can potentially be linked to multiple candidate Wikipedia pages but in this example only the mathematical reading is appropriate (`https://en.wikipedia.org/wiki/Number` and `https://nl.wikipedia.org/wiki/Getal_(wiskunde)`), and the same holds for the concept 'infinity' (linked to `https://en.wikipedia.org/wiki/Infinity`), 'oneindig' in Dutch, `https://nl.wikipedia.org/wiki/Oneindigheid`).

The annotation task was launched with an initial set of a 1,000 sentences for each language pair. In total, funding was available for 3,000 annotations of sentence pairs per language. As the contributors were also paid when annotating the test questions, which make up 30% of the total annotations, we could maximally expect to gather 700 items per language pair as we planned to use a redundancy of three in this task for quality purposes.

CrowdFlower allows for configuration settings that restrict the contributors according to several measures. Initially, we only allowed contributors that are residents of the countries in which the language was spoken. Unfortunately, the annotation process progressed very slowly for most of the languages and for some languages hardly any annotations were collected. Not only the availability of the contributors (limited by the restriction), but also the complexity of the task severely limited data annotation. The country limitation was therefore relaxed for all languages except PT for which we did have sufficient contributors. For Chinese specifically, we could not gather sufficient contributors via the CrowdFlower platform, and for this data set we asked a professional translator to annotate 200 sentences.

Lang pair		English			Target language		
	#s	slen	#t	#t/#s	slen	#t	#t/#s
BG	618	14.9	765	1.2	13.5	747	1.2
CS	688	14.4	712	1.2	11.7	702	1.2
DE	785	15.3	911	1.2	14.5	900	1.1
EL	627	23.0	923	1.2	22.9	912	1.3
HR	597	19.1	712	1.2	15.4	703	1.2
IT	782	22.3	1494	1.2	21.9	1477	1.3
NL	689	16.9	831	1.2	15.8	815	1.1
PL	569	16.2	653	1.3	13.5	638	1.3
PT	696	21.0	1289	1.3	20.2	1275	1.3
RU	837	17.4	1117	1.2	14.8	1088	1.2
ZH	604	19.6	856	1.2	14.0	759	1.1

Table 2: Data set statistics listing the number of sentence pairs (#s), the number of identified topics for both English and the target language (#t), average sentence length in tokens (slen), and average number of annotated topics per sentence (#t/#s).

5. Outcomes

At the end of the annotation process, we collected all annotations from the CrowdFlower platform and filtered the sentences to keep only those for which we had at least three trustworthy contributors and kept all annotations made by at least two annotators. This resulted in a data set in which for each language pair at least 500 sentences were annotated, as is shown in the second column in Table 2. When we compare the number of topics annotated in English (column 4 of Table 2) and in its corresponding topics in the target language (column 7), we can see that only for a handful of topics no equivalent phrase was available in the translation but in the vast majority the corresponding topic was marked. The average number of annotated topics per sentence is rather low, between 1.1 and 1.3 on average for all language pairs.

We also performed a manual inspection of the filtered trustworthy annotations to verify whether they are indeed correct, and to what extent they are complete. We took a small sample of 20 sentences for each language pair and counted the number of topics in English with their corresponding Wikipedia links. This analysis showed that annotations gathered via crowdsourcing were not complete; many topics that had a corresponding English Wikipedia page were not marked. For some languages even 50% of the topics that could have been marked were missing in this small sample. There are two main reasons for this. On the one hand, the annotators were not all consistent in their annotations and an entity would be marked by one annotators but not by all. As we only take into account those annotations made by at least two annotators to guard quality, we are losing these annotations. Also, many annotators chose to mark a minimum of entities instead of marking all options. However, this manual inspection showed us that the topics that were marked, were correctly marked, linked and also correctly linked to the topic in the target language (in this small sampling 90–100% correct). In terms of precision, we achieved a satisfying result.

The resulting data set will be made available through the EU (according to the H2020 Open Research Data Pilot) for research purposes after the end of the project, excluding the sentences sampled from Coursera as these were restricted by copyright to project-internal usage only. Therefore the publicly available data set contains around 40% less sentence pairs than the full data set shown in Table 2. The numbers of sentences in the publicly available dataset are shown in the second column in Table 3.

Table 3 represents the upper bound of Wikipedia coverage score by checking how many of the English manual labels had a corresponding Wikipedia page in the target language. This shows how many of the cases received a null value in the annotation process as discussed above. As expected, both Greek and Croatian have a low upper bound of around 60% indicating that due to the small number of Wikipedia pages for these languages, linking translations

lang	#s	upper bound
BG	461	76.23
HR	27	68.29
CS	529	82.31
DE	490	91.64
EL	467	62.20
IT	503	89.36
NL	482	87.10
PL	442	79.74
PT	529	92.95
RU	487	88.61
ZH	217	80.62

Table 3: The percentage of entities in English sentences that actually have an equivalent Wikipedia page in the target language.

lang	#s	precision
BG	461	56.21
HR	27	83.33
CS	529	57.63
DE	490	54.55
EL	467	57.14
IT	503	61.41
NL	482	52.17
PL	442	45.45
PT	529	64.41
RU	487	52.94
ZH	217	68.00

Table 4: Comparison between manual labels and labels produced by English Illinois Wikifier.

via Wikipedia is less informative for these languages than for the languages with better coverage.

6. Usage

We developed the dataset presented in this paper for the purpose of creating and tuning a method for implicit translation evaluation. We make use of the fact that most Wikipedia pages have translations in many other languages. We apply the following method. The input for evaluation is an English source sentence and its translation in the target language. We apply a Wikifier to find and link the entities in the English source data to their relevant Wikipedia pages. We check whether any entities identified in the source text have corresponding Wikipedia pages in the target languages using the inter-language links present in Wikipedia. Next we verify whether the target sentence indeed contain this corresponding translation, checking both for synonyms and lemmatized versions of the entity. When such a match is found, we count this as a correct entity translation, or as an error when no match was found.

The full details on the experiments for implicit evaluation are beyond the scope of this paper. To demonstrate the usability of the created data set we show the results of applying a state-of-the-art publicly available Wikifier on the English sentences in the data set. We apply the Illinois Wikifier (Cheng and Roth, 2013) that uses both global and local context cues (Ratinov et al., 2011) to disambiguate ambiguous names and concepts that have more than one possible Wikipedia page. The Wikifier first detects names and concepts in the text and then aims to link these entities to Wikipedia pages. In Table 4 we show a comparison between the manually labeled Wikipedia links and the links assigned by the Wikifier. The first column shows the number of sentences in the data set. We computed how many of the predicted Wikipedia links were correct (precision) shown in the second column. We observe that between 45 and 68% of the manual labeled entities was also marked by the Wikifier (the HR data set achieves 83% but it only contains 27 sentences). This is considerably lower than the outcome scores reported by (Cheng and Roth, 2013). When inspecting the mismatches between the Wikifier and the data set it becomes clear that the Wikifier, which is tuned on news articles, focuses mostly on named entities which are less present in the educational material. Typical educational entities like 'number' and 'infinity' from the example sentence in figure 1 are not recognized as entities by the Wikifier.

7. Conclusion and Discussion

In this article, we describe the process of creating a wikification data set. This dataset consists of pairs of sentences for eleven language pairs that are manually annotated with names and noun phrases. Each annotated item is linked to its corresponding Wikipedia page in the language of the sentence, if such a page exists. For the annotation process, we used a popular crowdsourcing platform to find contributors from the countries where these languages were spoken. The produced annotations are precise and correct, but lack in coverage (recall). We gave a brief example of how the data set can be used for the evaluation of translation or the evaluation of automatic wikification.

Malicious behavior in crowdsourcing platforms is a common phenomenon. Contributors are mainly interested in quick payment, making the least possible effort. The erroneously marked entities indicated that contributors paid little attention to the detailed instructions, which should be expected in a complex task, such as entity annotation. Furthermore, after a number of submissions, contributors get trained on the task process, and often adopt malicious behavior in order to cheat the system, despite the fact that we put in 30% test questions in the crowdsourcing setup. In a characteristic example, in a sentence with multiple entities, even decent contributors annotated only one, just to receive the reward easily and quickly.

The wikification annotation task is more complex than traditional crowdsourcing annotation tasks. Annotation of items in sentences should only be done if corresponding Wikipedia pages can be found, which means that an external source of information (Wikipedia) is needed to annotate the sentences. Additionally, after marking the items, the URL to the Wikipedia page should also be given. The same has to happen in the sentence in the other language of the language pair. This process consists of several complex

actions.

Given the complexity of the task, it is hard to find contributors who are interested in performing the task. Because the contributors also need to be fluent in the two languages of the language pair, this resulted in very limited availability of contributors for this task. Even when the initial country restriction was dropped, this still resulted in a small crowd for the task.

Ideally, the crowdsourcing platform opens a crowdsourcing job up to a large crowd all over the world. In practice, however, we did not easily get sufficient contributors for this specific task. We expect this is partly due to the complexity of the task in connection to the payment rate: we ended up with annotations by contributors from several low-income economies like Philippines, Venezuela, and Indonesia. For future entity annotation tasks, it would be worth considering using another crowdsourcing platform (e.g., MTurk) to target an a-priori competent pool of contributors, taking advantage of the platform features (e.g., qualified contributors with a specific success rate threshold in related past tasks).

We did consider reducing the complexity of the task by splitting the entire annotation process in several smaller steps, such as identification in English, identification in the other language, and linking to Wikipedia in both languages. Such approach is suggested by Kittur and colleagues (Kittur et al., 2011) who showed that a complex task can be broken down into smaller tasks. Such an alternative approach had several disadvantages. First, we would have to create multiple interfaces for the different types of tasks and setting up crowdsourcing jobs for each of the steps. After annotation, the annotations in the two languages would also need to be linked to each other. In the end, such a step-wise approach would have been much more costly and time consuming, while at the same time it is unlikely that it would increase annotation throughput. The most time consuming part of this annotation task was the step to look up each potential name and candidate in Wikipedia. If we were to split the task in multiple steps, this time consuming part would have to be performed multiple times.

While crowdsourcing is an excellent means of collecting annotated data for several purposes, it seems to be less suitable for complex tasks as the one described here. The task needs to be clear, straightforward, and annotation should not consist of several actions. Having strict requirements on the people in the crowd (such as country or language restrictions) severely limits the availability of the crowd, which in turn reduces the annotation speed.

From a financial point of view, crowdsourcing, even for complex tasks, allows for the collection of annotations that are otherwise simply not feasible. In terms of value for money, crowdsourcing leads to results that we could not have achieved in any other way. When attempting crowdsourcing data collection, care has to be taken on how to setup the tasks to maximise the use of the crowd, as well as find automated ways to ensure continuous monitoring of the crowd contributors' quality to ensure the successful completion of the task.

Acknowledgements

This work was completed as part of the TraMOOC project (Translation for Massive Open Online Courses) funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under grant agreement number 644333.

References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Agirre, E., Banea, C., Cer, D. M., Diab, M. T., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of SemEval@ NAACL-HLT*, pages 497–511.

Bontcheva, K., Derczynski, L., and Roberts, I. (2017). Crowdsourcing named entity recognition and entity linking corpora. In *Handbook of Linguistic Annotation*, pages 875–892. Springer.

Bougrine, S., Cherroun, H., and Abdelali, A. (2017). Altruistic crowdsourcing for arabic speech corpus annotation. *Procedia Computer Science*, 117:137–144.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.

Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232. Association for Computational Linguistics.

Cheng, X. and Roth, D. (2013). Relational inference for wikification. In *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1787–1796. Association for Computational Linguistics (ACL).

Cohen, K. B., Fort, K., Adda, G., Zhou, S., and Farri, D. (2016). Ethical issues in corpus linguistics and annotation: Pay per hit does not affect effective hourly rate for linguistic resource development on amazon mechanical turk. In *ETHics In Corpus collection, Annotation and Application workshop*.

Good, B. M., Nanis, M., Wu, C., and Su, A. I. (2014). Microtask crowdsourcing for disease mention annotation in pubmed abstracts. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 282–293. World Scientific.

Kittur, A., Smus, B., Khamkar, S., and Kraut, R. E. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 43–52, New York, NY, USA.

Liu, A., Soderland, S., Bragg, J., Lin, C. H., Ling, X., and Weld, D. S. (2016). Effective crowd annotation for relation extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906.

Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA. ACM.

Moro, A. and Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

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

Sabou, M., Bontcheva, K., Derczynski, L., and Scharl, A. (2014). Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of LREC*, pages 859–866.

Su, R., Shi, S., Zhao, M., and Huang, H. (2017). Utilizing crowdsourcing for the construction of chinese-mongolian speech corpus with evaluation mechanism. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 55–65. Springer.

Tsai, C.-T. and Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of HLT-NAACL*, pages 589–598.

van Gompel, M., van der Sloot, K., and van den Bosch, A., (2017). *Ucto: Unicode Tokeniser,Reference Guide, version 0.9.6*. Centre for Language Studies, Radboud University Nijmegen, January. URL: https://languagemachines.github.io/ucto/.

Wang, A., Hoang, C. D. V., and Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.