QTLeap WSD/NED Corpora:
Semantic Annotation of Parallel Corpora in Six Languages

Arantxa Otegi,[*] **Nora Aranberri**[*], **Antonio Branco**[‡], **Jan Hajič**[†], **Steven Neale**[‡],
Petya Osenova [▽], **Rita Pereira**[‡], **Martin Popel**[†], **João Silva**[‡], **Kiril Simov**[▽] **and Eneko Agirre**[*]
[*] University of the Basque Country (UPV/EHU), IXA Group – {arantza.otegi, nora.aranberri, e.agirre}@ehu.eus
[†] Charles University in Prague, Faculty of Mathematics and Physics, UFAL – {hajic, popel}@ufal.mff.cuni.cz
[‡]Universidade de Lisboa – {antonio.branco, steven.neale, ana.pereira, jsilva}@di.fc.ul.pt
[▽]Institute of Information and Communication Technologies (IICT-BAS) – {petya, kivs}@bultreebank.org

Abstract

This work presents parallel corpora automatically annotated with several NLP tools, including lemma and part-of-speech tagging, named-entity recognition and classification, named-entity disambiguation, word-sense disambiguation, and coreference. The corpora comprise both the well-known Europarl corpus and a domain-specific question-answer troubleshooting corpus on the IT domain. English is common in all parallel corpora, with translations in five languages, namely, Basque, Bulgarian, Czech, Portuguese and Spanish. We describe the annotated corpora and the tools used for annotation, as well as annotation statistics for each language. These new resources are freely available and will help research on semantic processing for machine translation and cross-lingual transfer.

Keywords: annotated parallel corpora, named-entity disambiguation, word sense disambiguation, coreference

1. Introduction

From a machine translation (MT) perspective, the deeper the processing of utterances, the less language-specific differences will remain between the representations of the meaning of the source and target texts. As a result, chances of success are expected to increase considerably by MT systems that are based on deeper semantic engineering approaches. Following this assumption, one of the approaches taken by the QTLeap project[1] is to enrich MT training resources with lexico-semantic information.

In this work, we present a solid effort to build multilingual parallel corpora annotated at multiple semantic levels. Our overall goal is to enrich two parallel corpora, Europarl (Koehn, 2005) and the QTLeap corpus (Agirre et al., 2015b), with token, lemma, part-of-speech (POS), named-entity recognition and classification (NERC), named-entity disambiguation (NED), word-sense disambiguation (WSD) and coreference for six languages covered in the QTLeap project, namely, Basque (EU), Bulgarian (BG), Czech (CS), English (EN), Portuguese (PT) and Spanish (ES). Specifically, this paper presents the first release of such corpora, which includes NERC, NED, WSD and coreference-level annotation for these six languages. Additionally, some languages have extra annotations, such as wikification (EN, ES), dependency parsing (BG, CS, EU) or constituency parsing (EN, ES), and semantic-roles (EN, ES). The annotated "Europarl-QTLeap WSD/NED corpus" and "QTLeap WSD/NED corpus" are distributed under the license CC BY-NC-SA 4.0, and have been released through Meta-share[2] and CLARIN Lindat.[3]

This paper is organized as follows: Section 2 presents the corpora we annotated; Section 3 describes the NERC, WSD, NED and coreference tools and annotation formats; Section 4 addresses the evaluation of the tools; Section 5 presents the corpus statistics of the annotation; and finally Section 6 outlines the conclusions.

2. Target Corpora

This section describes the two corpora we have annotated, the Europarl corpus and the QTLeap corpus. Each resource covers a different text-type and domain.

2.1. Europarl corpus

As a first resource, we have annotated the widely-used Europarl corpus[4] (release v7). It consists of texts extracted from the proceedings of the European Parliament, which include versions in 21 European languages. Compiled with machine translation in mind, matching items were extracted and labeled with corresponding document IDs. Then, sentence boundaries were identified and aligned (for further collection and processing information, see Koehn (2005)). The Europarl corpus consists of monolingual data as well as bilingual parallel data with English as pivot language. In our effort, we have annotated the BG, CS, ES and PT parts of the corpus separately while the EN side of the ES-EN language pair was used as pivot language to link all six languages.

Given that Europarl does not include Basque, we annotated an alternative publicly available Basque-English parallel corpus, the GNOME corpus (Tiedemann, 2012), which includes GNOME localization files.

2.2. QTLeap corpus

The QTLeap corpus consists of 4,000 pairs of questions and respective answers in the domain of IT troubleshooting

[1] `http://qtleap.eu`

[2] `http://metashare.metanet4u.eu/go2/europarl-qtleap-wsdned-corpus`
`http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus`

[3] `https://lindat.mff.cuni.cz`, namely:
`http://hdl.handle.net/11234/1-1476`
`http://hdl.handle.net/11234/1-1477`

[4] `http://www.statmt.org/europarl/`

for both hardware and software, distributed in four 1,000-pair batches (Gaudio et al., 2016). This material was collected using a real-life, commercial online support service via chat.

The QTLeap corpus is a unique resource in that it is a multilingual data set with parallel utterances in different languages (Basque, Bulgarian, Czech, Dutch, English, German, Portuguese and Spanish). This multilingual resource was obtained by translating the original Portuguese corpus into a pivot language, English, and this into the remaining six languages.

The current annotated corpus covers the first 2,000 sentences of the QTLeap corpus, which have been used to train the MT systems in the project.

3. Annotation Tools

In this section, we describe the NERC, NED, WSD and coreference tools used to annotate the corpora. We have chosen the tools based on their performance and their ease of use. We also describe the annotation formats.

3.1. Named-entity recognition and classification

Basque, English and Spanish ixa-pipe-nerc is a multilingual NERC tagger, part of IXA pipes (Agerri et al., 2014). Every model has been trained with the averaged Perceptron algorithm as described in Collins (2002) and as implemented in Apache OpenNLP. The datasets used for training the models are the following: Egunkaria dataset for Basque, a combination of Ontonotes 4.0, CoNLL 2003 and MUC 7 for English, and CoNLL 2002 for Spanish.

Bulgarian The Bulgarian NERC is a rule-based module. It uses a gazetteer with names categorized in four types: Person, Location, Organization, Other. The identification of new names is based on two factors – sure positions in the text and classifying contextual information, such as, titles for persons, types of geographical objects or organizations, etc. The disambiguation module uses simple unigram-based statistics.

Czech NameTag[5] is an open-source trainable tool for NERC (Straková et al., 2014). NameTag is distributed as a standalone tool or a library, along with pre-trained linguistic models. In the Czech model, entities are classified into two-level hierarchy of categories consisting of 42 fine-grained categories merged into 7 super-classes.

Portuguese LX-NER is a NERC tool that handles the following types of expressions: Numbers (Arabic, Decimal, Non-compliant, Roman, Cardinal, Fraction, Magnitude class, Measures (Currency, Time, Scientific units), Time (Date, Time periods, Time of the day) and Addresses) and name-based expressions (Persons, Organizations, Locations, Events, Works, Miscellaneous). The number-based component is built upon handcrafted regular expressions. It was developed and evaluated against a manually constructed test-suite including over 300 examples. The name-based component is based on Hidden Markov Models technology and was trained over a manually annotated corpus of approximately 208,000 words (Ferreira et al., 2007).

3.2. Named-entity disambiguation

Basque ixa-pipe-ned-ukb performs NED based on UKB, a graph-based WSD tool (see Section 3.3.). The Wikipedia graph built from the hyperlinks between Wikipedia articles is used for the processing. This tool was successfully used for English NED (Agirre et al., 2015a).

Bulgarian NED annotations follow the same approach as the Bulgarian disambiguation module (see Section 3.3.), but the DBpedia classes are used instead of WordNet. The ontological hierarchy of DBpedia determines the more general categories for DBpedia instances (City, Politician, etc.) as subclasses of Person, Location and Organization. For other kinds of instances it relies on the most general category provided by the classification of the instance according to DBpedia. Then the standard module is adapted to use the new categories. In case the selected categories in the annotation are not sufficient for disambiguating among DBpedia instance URIs, we store all of them in the annotation.

Czech During the preparation phase for NED Named Entities Linking table was created. Each row of that table consists of the lemmatized Czech Wikipedia article's title, Czech Wikipedia URL and English DBpedia URL, based on Czech Wikipedia dump (containing Czech titles and corresponding Wikipedia URLs) and English-Czech DBpedia dump (containing Czech labels and English DBpedia URLs). Additionally, lemmatization and tagging for each title was applied using MorphoDiTa (Straková et al., 2014). For each entity that is detected by NameTag, its form is lemmatized. Then we search the table for the occurrences of this lemmatized form. In case of ambiguity, the algorithm picks up the most "popular" article. The popularity of the article is computed using Wikipedia page-to-page link records, so the article with the highest number of reference links is preferred.

English and Spanish ixa-pipe-ned module performs the NED task based on DBpedia Spotlight (Daiber et al., 2013). Assuming that a DBpedia Spotlight REST server for a given language is locally running, the module performs the disambiguation for each entity detected by NERC module. It offers the "disambiguate" and "candidates" service endpoints. The former takes the spotted text input and it returns the DBpedia resource page for each entity. The later is similar to disambiguate, but returns a ranked list of candidates.

Portuguese The NED module for Portuguese, LX-NED, uses DBpedia Spotlight to find links to resources about entities identified in pre-processed input text. It creates a process to run a Portuguese extraction of DBpedia Spotlight on a local server, then takes an input text pre-processed with lemmas, Part of Speech tags and named entities using the LX-Suite (Branco and Silva, 2006) and converts it to the 'spotted' format understood by Spotlight. This spotted input text is then disambiguated using DBpedia Spotlight, returning among other information links to existing Portuguese DBpedia resource pages for each named entity discovered.

[5] http://ufal.mff.cuni.cz/nametag

3.3. Word-sense disambiguation

Basque, English and Spanish ixa-pipe-wsd-ukb is based on UKB, a collection of programs for performing graph-based WSD (Agirre and Soroa, 2009). It applies the so-called Personalized PageRank on a Lexical Knowledge Base (LKB) to rank the vertices of the LKB and thus perform disambiguation. WordNet 3.0 is the LKB used for this processing.

Bulgarian The basic version of Bulgarian WSD is implemented on the assumption of one sense per discourse and bigram statistics.

Czech Two different approaches were used for Czech WSD. The first approach based on the work of Dušek et al. (2015) focuses on verbal WSD. The second approach followed for the annotation is a straightforward way of achieving compatibility with English WordNet IDs. Since the Czech corpus contains the same sentences as the English corpus, the English WordNet ID annotation from this corpus is projected onto Czech words using GIZA++ word alignment.

Portuguese The Portuguese WSD tool, LX-WSD, is also based on UKB. The LKB from which UKB returns word senses within the pipeline has been generated from an extraction of the Portuguese MultiWordNet[6].

3.4. Coreference

Basque ixa-pipe-coref-eu is an adaptation of the Stanford Deterministic Coreference Resolution (Lee et al., 2013), which gives state-of-the art performance for English. The original system applies a succession of ten independent deterministic coreference models or sieves. During the adaptation process, firstly, a baseline system has been created which receives as input texts processed by Basque analysis tools and uses specifically adapted static lists to identify language dependent features like gender, animacy or number. Afterwards, improvements over the baseline system have been applied, adapting and replacing some of the original sieves (Soraluze et al., 2015), taking into account that morphosyntactic features are crucial in the design of the sieves for agglutinative languages like Basque.

Bulgarian A basic version of a coreference resolution module uses paths in the dependency tree of each sentence. By using path patterns, anaphora resolution is mainly performed. When dealing with the rest of the word forms, the open class words that belong to the same synsets in WordNet are considered and grouped them together.

Czech There are multiple modules for Czech coreference, each of them aiming at a specific type of coreference: coreference of reflexive pronouns, relative pronouns, zeros, personal and possessive pronouns in 3rd person and coreference of noun phrases (Bojar et al., 2012; Novák and Žabokrtský, 2011). Coreference relations are annotated between the nodes of dependency trees that serve as a deep syntax representation of sentences. This enables the system to take advantage of rich linguistic annotations available in the trees as well as to resolve coreference even for subject pronouns dropped from the surface representation (zeros), which is a common practice in Czech.

English and Spanish ixa-pipe-coref is loosely based on the Stanford Multi Sieve Pass system (Lee et al., 2013). The system consists of a number of rule-based sieves. Each sieve pass is applied in a deterministic manner, reusing the information generated by the previous sieve and the mention processing. The order in which the sieves are applied favors a highest precision approach and aims at improving the recall with the subsequent application of each of the sieve passes. This is illustrated by the evaluation results of the CoNLL 2011 Coreference Evaluation task (Lee et al., 2013; Lee et al., 2011), in which the Stanford system obtained the best results. The results show a pattern which has also been shown in other results reported with other evaluation sets (Raghunathan et al., 2010), namely, the fact that a large part of the performance of the multi-pass sieve system is based on a set of significant sieves. Thus, this module so far focuses on a subset of sieves only, namely, Speaker Match, Exact Match, Precise Constructs, Strict Head Match and Pronoun Match (Lee et al., 2013).

Portuguese For the Portuguese coreference tool, a decision tree classifier was experimented with. Given a pair of expressions, the classifier returns a true or false value that indicates whether those expressions are coreferent. The classifier was trained over the Summit Corpus (Collovini et al., 2007) using the J48 algorithm in the Weka machine-learning toolkit. The most relevant features, according to the work of de Souza et al. (2008), were extracted from Summ-It and used to train the J48 algorithm with default parameters. The resulting decision tree produced by J48 turned up to very simple and boils down to comparing the cores and the morphological information (gender and number) of the two expressions. As such, we found it easier to directly implement equivalent tests in-code instead of having to feed the extracted features to the Weka J48 classifier proper.

3.5. Annotation formats

Basque, Bulgarian, Czech, English and Spanish These corpora are annotated in the NAF format. The NAF format (Fokkens et al., 2014) is a linguistic annotation format designed for complex NLP pipelines that combines strengths of the Linguistic Annotation Framework (LAF) and the NLP Interchange Formats described by Ide and Romary (2003). Because of its layered extensible format, it can easily be incorporated in a variety of NLP modules that may require different linguistic information as their input.

Portuguese The corpus for Portuguese is divided into 4 text files - the raw corpus, and one file for the output of each of the three tools used to process it (WSD, NED and coreference). For each of the three tools output is provided in a standoff annotation format, consisting of one token per line (ID of each token in a markable pair in the case of the coreference tool), the appropriate output element of the respective tools (word sense, named entity URI or true or false in the case of coreference), and additional metadata such as token IDs, sentence IDs and part-of-speech (POS) tags in the case of the WSD and NED tools.

[6] `http://multiwordnet.fbk.eu/english/home.php`

Language	NERC	NED	WSD	Coreference
Basque	76.72	87.90	56.40	53.67
Bulgarian	79.13	46.88	65.85	31.11
Czech	80.30	–	80.47	see 4.1.4.
English	86.21	77.76	80.10	56.40
Portuguese	85.73	67.07	65.00	–
Spanish	80.16	65.11	79.30	51.38

Table 1: F-scores for annotation tools. Note that evaluation sets vary across languages.

4. Evaluation

4.1. Evaluation on standard datasets

We provide the performance measurements of the tools used to annotated the QTLeap WSD/NED corpora, providing information on the evaluation datasets and scores for each tool. The summary performance as measured on standard datasets for all languages is presented in Table 1.

4.1.1. NERC

Basque A subset of the EPEC corpus including 60,000 tokens was manually annotated with 4,748 named entities.[7] When evaluated over a subset of ca. 15,000 tokens, ixa-pipe-nerc's F-score measure is 76.72% on 3 class evaluation and 75.40% on 4 classes.

Bulgarian The Bulgarian NERC tool was evaluated on a dataset of the BulTreeBank of 12,223 tokens.[8] The gold standard annotation contains 810 named entities. The F-score of the tool is 79.13%.

Czech NameTag is the state-of-the-art NERC tool for Czech. Its F-score on the test portion of Czech Named Entity Corpus 2.0[9] is 80.30% for the coarse-grained 7-classes classification and 77.22% for the fine-grained 42-classes classification (Straková et al., 2014).

English The ixa-pipe-nerc module based on CoNLL 2002[10] and 2003,[11] trained on local features only obtains F-score of 84.53%, and the models with external knowledge 87.11%. The OntoNotes CoNLL with 4 NE types and local features model obtains a F-score of 86.21%.

Portuguese The rule-based component of the NERC was evaluated against a manually constructed test-suite including over 300 examples. It scored 85.55% F-score. When trained over a manually annotated corpus of approximately 208,000 words and evaluated against an unseen portion with approximately 52,000 words, the data-based module scored a 85.73% F-score (Ferreira et al., 2007).

Spanish The ixa-pipe-nerc module for Spanish currently obtains the best results when training Maximum Entropy models on the CoNLL 2002 dataset. Our best model obtains a 80.16% F-score. The best result so far on this dataset is a 81.39% F-score (Carreras et al., 2002) when using external knowledge, and a 79.28% F-score without it.

4.1.2. NED

Basque The ixa-pipe-ned-ukb module for Basque was evaluated on the publicly available EDIEC (Basque Disambiguated Named Entities Corpus) dataset.[12] This dataset is a corpus of 1,032 text documents with manually disambiguated NEs (Fernandez et al., 2011). We obtained a performance of 87.90% in F-score (Pérez de Viñaspre, 2015).

Bulgarian The NED gold standard from the BulTreeBank-DB includes 667 instances annotated with DBpedia.[13] The F-score of the tool is 46.88%. The low results are due to the small coverage of the Bulgarian DBpedia.

Czech There is no publicly available Czech test set for NED, so we only performed a qualitative evaluation of the tool (see Section 4.2.2.).

English The ixa-pipe-ned module was evaluated on the TAC KBP 2011 dataset[14] and the AIDA corpus,[15] discarding NIL instances. The best results were a 68.93% F-score for TAC and a 77.76% F-score for AIDA.

Portuguese LX-NED was evaluated using the NE-annotated version of the CINTIL International Corpus of Portuguese (Barreto et al., 2006). Out of the 26,371 NEs in the CINTIL corpus, 16,120 were manually disambiguated using DBpedia. 12,160 of these were also automatically disambiguated by LX-NED. We thus define recall as the number of entities with the same DBpedia entry assigned by both the NED tool and the human annotator, divided by the number of entities manually disambiguated (16,120). The F-score for LX-NED is 67.07%.

Spanish The Spanish ixa-pipe-ned module was evaluated on the TAC 2012 Spanish dataset.[16] The system identifies entities on a Spanish document and links them to an English Knowledge Base using the interlingual links from Wikipedia.[17] We obtained a performance of 65.11% in F-score.

4.1.3. WSD

Basque The ixa-pipe-wsd-ukb module for Basque was evaluated on the publicly available EPEC-EuSemcor dataset.[18] This dataset is a Basque SemCor corpus, that is, a

[7] http://ixa2.si.ehu.es/eiec/eiec_v1.0.tgz

[8] http://www.bultreebank.org/dpbtb/

[9] http://ufal.mff.cuni.cz/cnec

[10] http://www.clips.ua.ac.be/conll2002/ner/

[11] http://www.clips.ua.ac.be/conll2003/ner/

[12] http://ixa2.si.ehu.es/ediec/ediec_v1.0.tgz

[13] http://www.bultreebank.org/QTLeap/

[14] Text Analysis Conference (TAC) for the Knowledge Base Population (KBP) track: https://www.ldc.upenn.edu/collaborations/current-projects/tac-kbp Datasets available on https://catalog.ldc.upenn.edu/

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

[16] Text Analysis Conference (TAC) for the Knowledge Base Population (KBP) track: https://www.ldc.upenn.edu/collaborations/current-projects/tac-kbp Datasets available at https://catalog.ldc.upenn.edu/

[17] http://www.mediawiki.org/wiki/Interlanguage_links

[18] http://ixa2.si.ehu.es/mcr/EuSemcor.v1.0/EuSemcor_v1.0.tgz

Basque sense-tagged corpus, annotated with Basque WordNet v1.6 senses (Pociello et al., 2011). More specifically, it contains 42,615 occurrences of nouns manually annotated, corresponding to the 407 most frequent Basque nouns. We obtained a 56.4% F-score.

Bulgarian The BulTreeBank-DB WSD gold standard includes 3,118 sense annotations. The tool obtained a 65.85% F-score. The result is relatively good, considering the limited size of the Bulgarian WordNet, used in the annotation.

Czech The verbal WSD approach (Dušek et al., 2015) was evaluated on the Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012) and showed a 80.47% F-score. For the second approach (Czech text annotated with English WordNet 3.0 IDs), there is no publicly available test set.

English The ixa-pipe-wsd-ukb module for English was evaluated on the general domain coarse-grained all-words datasets (S07CG) (Navigli et al., 2007). This dataset uses coarse-grained senses which group WordNet 2.1 senses. The WSD system was run using WordNet 2.1 relations and senses and the mapping from WordNet 2.1 senses. In order to return coarse grained-senses, the algorithm was run on fine-grained senses, and the scores were aggregated for all senses that mapped to the same coarse-grained sense. Finally, the coarse-grained sense with the highest score was chosen. The overall result obtained was a 80.1% F-score.

Portuguese LX-WSD was evaluated using the sense-annotated version of the CINTIL International Corpus of Portuguese (Barreto et al., 2006), manually annotated with ILIs from the Portuguese MultiWordNet[19] (approximately 19,700 verified synsets). We thus define recall as the number of words with the same sense assigned by UKB and the human annotator, divided by the number of words manually disambiguated (45,502). LX-WSD scored a 65.00% F-score.

Spanish The ixa-pipe-wsd-ukb module for Spanish was evaluated on SemEval-2007 Task 09 dataset (Màrquez et al., 2007). The dataset contains examples of the 150 most frequent nouns in the CESS-ECE corpus, manually annotated with Spanish WordNet synsets. We ran the experiment over the test part of the dataset (792 instances) and obtained a 79.3% F-score.

4.1.4. Coreference

Basque The ixa-pipe-coref-eu module has been evaluated on the publicly available EPEC-KORREF dataset.[20] This dataset is a corpus of Basque text documents with manually annotated mentions and coreference chains, which consists of 46,383 words that correspond to 12,792 mentions. Our best system scored 53.67% CoNLL F-score[21], 5 points above baseline (48.67% F-score) (Soraluze et al., 2015).

Bulgarian A part of BulTreeBank-DB dataset was annotated again by hand for coreference chains (1468 words, 37 coreference chains, including 154 phrases). The system returns 49 coreference chains, including 182 phrases. We have calculated the precision and recall as proposed in (Vilain et al., 1995). The measured F-score was 31.11%.

Czech Coreference resolvers for Czech were evaluated separately for three different classes of anaphors: relative pronouns, a joint group of subject zeros, personal, and possessive pronouns (all in 3rd person), and noun phrases. For each anaphor class, F-scores of finding any of its antecedents were measured on the evaluation set of Prague Dependency Treebank 3.0 (Bejček et al., 2013). The relative pronoun resolver, using rule-based approach, obtained a 67.04% F-score. The other two classes obtained a 50.28% F-score and a 44.40% F-score respectively, using a machine learning approach. The F-score is calculated from the counts of how often any of the anaphor's antecedent is found, collected over each of the anaphors individually. This evaluation approach is similar to the one presented by Tuggener (2014).

English The ixa-pipe-coref module was evaluated on the development auto section of the CoNLL 2011 shared evaluation task,[22] which uses the English language portion of the OntoNotes 4.0 corpus. It scored a 56.4% CoNLL F-score, around 3 points below Stanford's system.

Portuguese The Portuguese coreference tool was trained using the Summ-it Corpus (v3.0) (Collovini et al., 2007). For 316,000 sentences of the Portuguese side of Europarl (∼10 million tokens), the Portuguese Coreference tool was able to find 727,142 markable pairs, from which 22,984 (3.16%) are coreferent. One possible cause for the tool's low recall of markable pairs could be inconsistencies between the dependency-parsed and constituency-parsed inputs over which the tool runs.

Spanish The ixa-pipe-coref module for Spanish was evaluated on the publicly available datasets distributed by the SemEval 2010 task on Multilingual Coreference resolution, in which the AnCora-ES (the Spanish part) corpus is used, yielding a CoNLL F1 score of 63.4.

4.2. Qualitative evaluation on the QTLeap in-domain corpus

In this section, we include a qualitative evaluation of the tools when applied to the IT-domain QTLeap corpus. We give results per tool type, as conclusions for all languages were very similar.

4.2.1. NERC

The NERC tools remained at high accuracy level, although we identified a drop in performance due to change of domain. First, we observed a drop in recall. IT-domain texts include a considerable amount of brand and product names and the tools are not trained to identify them. As an example, out of the 85 occurrences of *Skype*, the Czech tool recognizes 15. Secondly, precision decreased by frequent

[19] http://multiwordnet.fbk.eu/english/home.php

[20] http://ixa2.si.ehu.es/epec-koref/epec-koref_v1.0.tgz

[21] The CoNLL F-score is the average of the MUC, CEAF and B-CUBED F-scores (Pradhan et al., 2011)

[22] http://conll.cemantix.org/2011/introduction.html

Corpus	Basque		Bulgarian		Czech		English		Portuguese		Spanish	
tokens	53,239		67,591		71,061		68,913		72,018		71,989	
in WordNet	24,691	46.38%	12,627	18.7%	11,060	15.5%	25,807	37.45%	6,116	20.40%	22,704	31.54%
entities	869		180		1715		2,999		3,799		4,313	
in DBpedia	252	29.00%	180	100%	572	33.3%	1,950	65.02%	1,868	49.17%	3,175	73.61%
coreference chains	5,542		306		1,027		1,199		183		705	

Table 2: Statistics on the QTLeap WSD/NED corpus for six languages.

Corpus	Basque		Bulgarian		Czech		English		Portuguese		Spanish	
tokens	5.16M		4.84M		9.09M		52.24M		5.04M		57.00M	
in WordNet	2.21M	42.94%	1.28M	26.50%	4.47M	49.20%	22.70M	43.46%	0.66M	32.13%	20.20M	35.43%
entities	0.07M		0.06M		0.30M		1.91M		0.17M		2.18M	
in DBpedia	0.03M	40.26%	0.15M	39.40%	0.12M	39.40%	1.50M	78.24%	0.10M	59.58%	1.21M	55.49%
coreference chains	0.72M		0.03M		0.20M		1.25M		5.88M		0.94M	

Table 3: Statistics on the Europarl-QTLeap WSD/NED corpus for six languages.

mislabeling for IT-specific entities. *USB*, *Wi-Fi* and *Internet* are all classified as Organization by the Basque tool. Similarly, *Windows*, *Facebook* and *Google* are often classified as Location by the Spanish, English and Portuguese tools. The mislabeling is exacerbated by the inherent ambiguity between company and product names. The analyses showed that classification modules are not tuned to deal with terminology, product names or highly instructive text, which is a known weakness of NERC tools trained on general corpora.

4.2.2. NED

The domain-specific entities identified by the NERC tools were mostly correctly linked to DBPedia, and therefore, the disambiguation tools seem to perform at the expected level. For instance, for Basque *Sareko* (net) and *Facebook* were linked to `http://eu.dbpedia.org/resource/Internet` and `http://eu.dbpedia.org/resource/Facebook`, respectively. Even domain-specific products such as *Java* and *MB* were correctly linked to `http://eu.dbpedia.org/resource/Java_(programazio_lengoaia)` and `http://eu.dbpedia.org/resource/Megabyte`. We see, however, some room for improvement. Firstly, in-domain terminology still poses some difficulty, with cases such as *PC*, incorrectly linked to Microsoft Windows by the Czech tool `http://cs.dbpedia.org/resource/Microsoft_Windows`, or *PS* which was incorrectly linked to the French Socialist Party `http://eu.dbpedia.org/resource/Frantziako_Alderdi_Sozialista` by the Basque tool, when it was referring to PlayStation console. Secondly, the incorrect cases due to the lack of shared Wikipedia/DBpedia entries for the working languages is notable.

4.2.3. WSD

Word-sense disambiguation was based on WordNet for Basque, English, Portuguese and Spanish, and on Valency Lexicon (Urešová, 2011) for Czech. WSD performance was reasonable in the IT-domain, with little loss in accuracy. The decrease was mainly due to missing terms in the WordNets of the languages in question or incorrect assignments of synsets/valencies. Such is the case of the domain-specific *banda*, for instance, which was linked to the synset 30-04339291 with a confidence of 0.219025, referring to an *artifact consisting of a narrow flat piece of material*, instead of the correct synset 30-06260628, which is the specific synset for the domain of telecommunications *a band of adjacent radio frequencies (e.g., assigned for transmitting radio or television signals).*

4.2.4. Coreference

The QTLeap corpus is quite peculiar from a coreference point of view. The user-machine interactions generally consist of one user question and one answer. The answer usually consists of one sentence, but occasionally a few short sentences are displayed. In this context, the number of coreference chains present in the texts is low. For example, the Czech resolvers found only a very small number of coreferent pairs (1,860 from 82,496 markable pairs).

5. Statistics

The statistics for the annotated corpora are shown in Tables 2 and 3. They report the number of tokens of the texts included in each corpus, the entities found by each of the named entity recognition tools and the annotated coreference chains. Additionally, the number of terms linked to WordNet by the word sense disambiguation tools and the number of entities linked to DBpedia by the named entity disambiguation tools are also displayed.

6. Conclusions

This paper presents two multilingual parallel corpora automatically annotated with lexico-semantic information for six languages. The corpora comprise both the well-known corpus and a domain-specific question-answer troubleshooting corpus in the IT domain. This release includes NERC, NED, WSD and coreference-level annotation for Basque, Bulgarian, Czech, English, Portuguese and Spanish. We include references to the tools used, as well as the evaluation of each tool on standard data sets, and a qualitative assessment on the domain-specific QTLeap corpus.

The primary goal of this effort is to enrich machine translation training resources with cross-lingual lexico-semantic information. This information helps abstract linguistic forms and reduce source and target language differences during translation, increasing the probabilities of success. Additionally, however, this resource can be useful for cross-lingual transfer. This resource is publicly available through Meta-share and CLARIN Lindat.

Acknowledgments

This work has received support by the EC's FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches". This work has been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Agerri, R., Bermudez, J., and Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41. Association for Computational Linguistics.

Agirre, E., Barrena, A., and Soroa, A. (2015a). Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation.

Agirre, E., Branco, A., Popel, M., and Simov, K. (2015b). Europarl QTLeap WSD/NED corpus. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M. F. B., Nunes, F., and Silva, J. (2006). Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC '06, pages 1438–1443.

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague Dependency Treebank 3.0.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., and Tamchyna, A. (2012). The joy of parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey. European Language Resources Association.

Branco, A. and Silva, J. (2006). A Suite of Shallow Processing Tools for Portuguese: LX-Suite. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations*, EACL '06, pages 179–182. Association for Computational Linguistics.

Carreras, X., Màrquez, L., and Padró, L. (2002). Named Entity Extraction Using AdaBoost. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Collovini, S., Carbonel, T. I., Fuchs, J. T., Coelho, J. C., Rino, L., and Vieira, R. (2007). Summ-it: um corpus anotado com informações discursivas visando sumarização automática. In *Proceedings of TIL 2007*.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124. ACM.

de Souza, J. G. C., Gonalves, P. N., and Vieira, R. (2008). Learning Coreference Resolution for Portuguese Texts. In A. Teixeira, et al., editors, *Computational Processing of the Portuguese Language*, volume 5190 of *Lecture Notes in Computer Science*, pages 153–162. Springer Berlin Heidelberg.

Dušek, O., Fučíková, E., Hajič, J., Popel, M., Šindlerová, J., and Urešová, Z. (2015). Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation. In Eva Hajičová et al., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 82–90, Uppsala, Sweden. Uppsala University, Uppsala University.

Fernandez, I., Alegria, I. n., and Ezeiza, N. (2011). Semantic Relatedness for Named Entity Disambiguation Using a Small Wikipedia. In Ivan Habernal et al., editors, *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 276–283. Springer Berlin Heidelberg.

Ferreira, E., Balsa, J., and Branco, A. (2007). Combining rule-based and statistical methods for named entity recognition in Portuguese. In *In V Workshop em Tecnologia da Informao e da Linguagem Humana*, pages 1615–1624.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., van Hage, W. R., and Vossen, P. (2014). NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9.

Gaudio, R., Branco, A., and Burchardt, A. (2016). QTLeap – A Corpus for Cross-Lingual Information Dialogue. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association.

Ide, N. and Romary, L. (2003). Outline of the international standard linguistic annotation framework. In *Proceedings of the ACL 2003 workshop on Linguistic annotation: getting the model right-Volume 19*, pages 1–5. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Màrquez, L., Villarejo, L., Martí, M. A., and Taulé, M. (2007). SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 30–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Novák, M. and Žabokrtský, Z. (2011). Resolving Noun Phrase Coreference in Czech. *Lecture Notes in Computer Science*, 7099:24–34.

Pérez de Viñaspre, J. (2015). Wikipedia eta anbiguetate lexikala. Technical report, Computer Science Faculty, University of the Basque Country.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Soraluze, A., Arregi, O., Arregi, X., and de Ilarraza, A. D. (2015). Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. *Procesamiento del Lenguaje Natural*, 55:23–30.

Straková, J., Straka, M., and Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC*, pages 2214–2218.

Tuggener, D. (2014). Coreference Resolution Evaluation for Higher Level Applications. In Gosse Bouma et al., editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 231–235. The Association for Computer Linguistics.

Urešová, Z. (2011). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Prague.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, pages 45–52.