Named Entities in the QTLeap Corpus of Online Helpdesk Interactions

Andreia Querido, Rita de Carvalho, João Rodrigues, João Silva, Steven Neale, Rita Valadas Pereira, Patrícia Gomes, Catarina Correia, Diana Amaral, António Branco[1]

Faculdade de Ciências da Universidade de Lisboa

Abstract:
In this paper we present the annotation of a corpus with named entities that are classified into semantic types and disambiguated by linking them to their corresponding entry in the Portuguese DBpedia. This corpus, QTLeap Corpus, is a multilingual collection of question and answer pairs from a chat-based helpdesk service for Information and Communication Technologies. The resulting annotated corpus is a gold-standard named entity annotated lexical resource that is useful in supporting the training and evaluation of named entity annotation and disambiguation tools for Portuguese.

Keywords: annotated corpus, QTLeap Corpus, named entities, annotation task, disambiguation task.
Palavras-chave: corpus anotado, Corpus QTLeap, expressões de nomeação de entidades, tarefa de anotação, tarefa de desambiguação.

1. Introduction

This paper presents the annotation of QTLeap Corpus with classified and disambiguated named entities, and describes the methodology that guided this annotation. The innovative aspects of this work are in the genre of the corpus and in the information that the corpus is annotated with.

[1] {andreia.querido; rita.carvalho; joao.rodrigues; jsilva; steven.neale; ana.pereira; patricia.gomes; catarina.correia; diana.amaral; antonio.branco}@di.fc.ul.pt

The QTLeap Corpus (Del Gaudio *et al.*, 2015) consists of a collection of real interactions, carried out through a textual chat channel, between clients and professionals of a helpdesk service for Information and Communication Technologies (ICT). The corpus is multilingual and aligned, having been translated into seven other languages. The present paper addresses the named entity annotation of the Portuguese corpus.

The named entity annotation of QTLeap Corpus was done through two distinct and complementary tasks. The first task is manual named entity recognition and classification, or NERC, where named entities are marked and classified into specific semantic categories (like Organization, Time or Person). The second task is manual named entity disambiguation, or NED, in which the referent of each NE is disambiguated by linking the NE to the corresponding entry in the Portuguese DBpedia.

This dataset has the potential to present interesting features to study named entities, as will be shown ahead. It also contributes a gold-standard resource of classified and disambiguated named entities in Portuguese, that can support the development of NERC and NED tools for Portuguese either as dedicated training materials or as gold data for evaluating tools.

2. The QTLeap Corpus

The QTLeap Corpus was created in the scope of the QTLeap project[2] – Quality Translation by Deep Language Engineering Approaches –, a project whose goal is to develop a novel methodology for automatic translation that makes use of deep linguistic analysis to improve translation quality. In the context of Machine Translation, named entities (NE) need a specific analysis since they often cause translation failures, for instance when translating a named entity as a common noun (Babych and Hartley, 2003).

The QTLeap Corpus can be categorized as belonging to the Information and Communication Technologies (ICT) domain and is composed of linguistic interactions between users and providers of an ICT helpdesk service. These interactions are provided by a Portuguese company, Higher Functions, that ensures technical support to its clients, helping them solve problems

[2] Mais informações em QTLeap: http://qtleap.eu

related to software and hardware. These interactions are in the form of written questions made by users and the corresponding written answers provided by professionals working in the helpdesk. The corpus, originally in Portuguese, was translated into the seven languages of the partners involved in the QTLeap project, namely Spanish, Czech, German, English, Bulgarian, Basque and Dutch. As a result, QTLeap Corpus is also an aligned multilingual corpus. The corpus is a collection of 4 000 question-answer pairs which, in the Portuguese corpus, amount to 123 982 tokens. Currently, ¾ of the corpus has been annotated and disambiguated concerning named entities, for a total of 68 928 tokens over 3 000 question-answer pairs.

The annotation of this corpus is crucial to train and evaluate tools and components in the scope of the project. Besides that, it is also useful because it represents a domain and a genre that is seldom studied, even though ICT is a growing and dynamic area. To the best of our knowledge, there is no annotated corpus for Portuguese with these characteristics.

3. Named entity recognition and classification task

3.1 Methodology

3.1.1 Annotation task

The task of named entity recognition and classification (NERC) consists of identifying the NEs in the corpus. As the NEs are identified, they are also classified by assigning to each NE one of the ten pre-established semantic types, which are presented below.

To ensure a reliable linguistically interpreted dataset, manual annotation is done by two annotators working under a double-blind scheme followed by a phase of data curation where a third annotator adjudicates any mismatches between the two annotators. All the annotators have graduate or postgraduate education in Linguistics or similar fields.

The NEs were annotated with the semantic types presented in Barreto *et al*. (2006). The list of these semantic types, illustrated with examples taken from the corpus, may be found below:

Numbers (NUM)

(1) “No canto inferior direito do écran tem o símbolo do wireless (**5** traços) (...)”

"In the lower right corner of the screen it is the wireless symbol (**5** bars) (...)"

Measure (MEA):

(2) "Desligue o router da corrente eléctrica e volte a ligar passados **30 segundos**."

"Turn off the router of the electric power and turn on again, after **30 seconds**."

Time (TIME):

(3) "Como se trata de uma versão bastante antiga (ano de **2001**) à partida não irá funcionar."

"Given it is an old version (year of **2001**), it will probably not work."

Addresses (ADR):

(4) "Tente verificar no site (**https://drive.google.pt**) se ele se encontra no separador Lixo."

"Try checking in the site (**https://drive.google.pt**) if it is in Trash tab."

Persons (PER):

No occurrences in the corpus.

Organizations (ORG):

(5) "Utilize as mesmas credenciais de login na área de cliente **PT**."

"Use the same login credentials in **PT** customer's area."

Locations (LOC):

(6) "Porque é que não consigo abrir o facebook na **China**?"

"Why cannot I open facebook in **China**?"

Events (EVT):

No occurrences in the corpus.

Works (WRK):

No occurrences in the corpus.

Miscellaneous (MSC):

(7) "Por exemplo o nome do ficheiro é **teste.txt.exe**"

"For example, the name of file is **teste.txt.exe**"

(8) "Um **ID Apple** é o seu nome de utilizador para todas as suas interacções com a Apple"

"The **Apple ID** is your username for all your interactions with Apple"

After this first stage of annotation, we realized that we could not proceed without a more detailed set of semantic types for NEs as there was a high percentage (90.65%) of NEs that were placed in the category of Miscellaneous (see Figure 1), since, due to the specificity of the corpus, none of the other existing categories was suitable. Therefore, the corpus was re-annotated using a more fine-grained categorization and tagset.

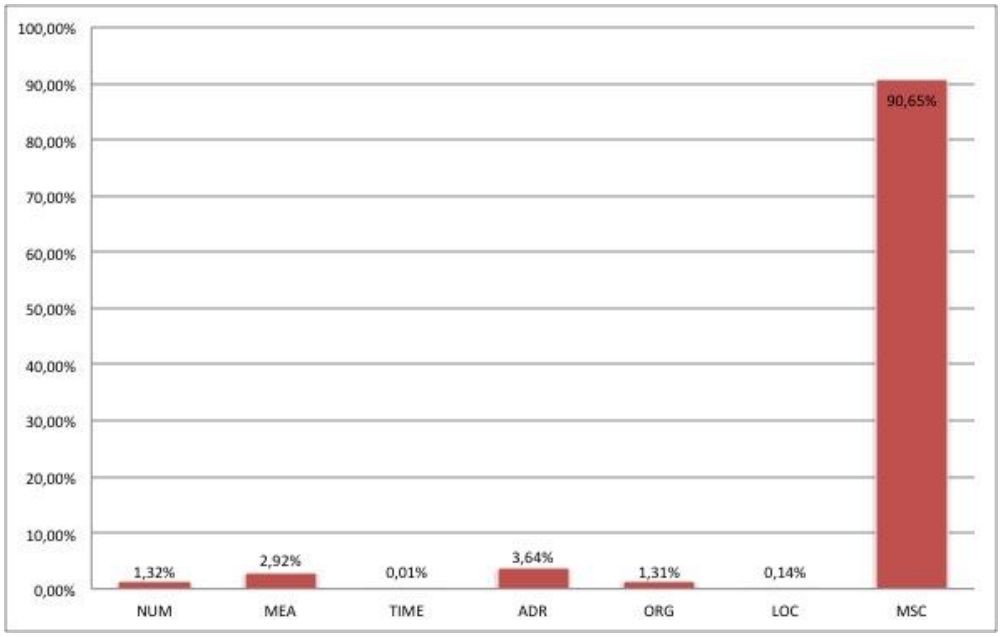

Figure 1: Distribution of NEs (first stage)

According to the available data and the results, three new domain-specific semantic types were created. Below follows these semantic types with examples taken from the corpus:

Software and Services (SS):

(9) "Porque é que não consigo abrir o **facebook** na China?"

"Why cannot I open **facebook** in China?"

(10) "Como desligar o **Windows 7**"

"How to turn off **Windows 7**"

Commands (CMD):

(11) "No menu inserir seleccione **Imagem**"

"In the menu insert select **Image**"

(12) "Clique na tecla **F5**"

"Press the **F5** key"

Hardware (HDW):

(13) "Acabo de comprar um novo **HP Probook** e costumo colocá-lo no meu colo"

"I just bought a new **HP Probook** and I usually put it in my lap"

(14) "O **Ipad** não liga."

"The **Ipad** does not turn on."

The distribution of NEs after re-annotation with a set of semantic types that includes these three new categories will be presented in Section 3.3.

3.1.2 Annotation tool

WebAnno is a general-purpose web-based annotation system (Yiman *et al.*, 2014). Despite being under development, it is quite stable and already possesses a set of features that are useful for the annotation we need to carry out: WebAnno allows us to create an annotation project and

fully customize it by specifying each annotation layer in terms of its set of valid tags and type. In this case, one annotation layer was created with a set of tags that comprises each type of NE.

WebAnno allows us to import files in plain text format (txt) and to export the annotated text in several formats (plain text, binary format, XMI format, WebAnno TSV format and Weblicht TCF format).

The annotation process itself is supported by a user-friendly and intuitive interface that allows editing a tag by clicking on it and connecting discontinuous NEs by dragging an arc between them. Being web-based, WebAnno runs directly in the browser, which means that is not necessary to install any specific software on the computers used by the annotators and that all annotated files are automatically stored in the server. This is coupled with a project management feature that allows for the administrator of a project to distribute the files to be annotated among the annotators, and a curation feature that automatically finds mismatches between annotators.

3.2 Challenges

Identifying NE semantic types requires a special attention to the context, hence "although they are individualizing labels, proper nouns, in general, are not unique tags that can only be applied to one single entity in the world" (Paiva Raposo & Bacelar do Nascimento, 2013)[3]. That is, the information that the context gives us is crucial for the identification of the referent of each NE (*idem, ibidem*).

The fact that the annotation was done sentence by sentence, in groups of documents in which questions and answers were separated, was, undoubtedly, one of the major annotation challenges since the annotator only had access to the immediate context, i.e., only the question or the answer in which the NE occurs and not the complete question-answer pair. Because of this, the classification of some NEs was sometimes difficult. When in doubt, these NEs were always marked as "MSC".

3 Our translation of "embora sejam etiquetas individualizadoras, os nomes próprios, em geral, não são etiquetas únicas que se apliquem a uma só entidade do mundo" (Paiva Raposo & Bacelar do Nascimento, 2013: 996).

More so than the limited context that the annotator could access, the annotation of such a domain-specific text raised questions that can only be overcome by a very specific knowledge of the world. Recurrently, to be able to determine the semantic type of a NE, or, even before that, to be able to identify whether an expression was a NE, knowledge of the ICT domain was essential to understand the meaning of the expression in that particular context.

For example, in a sentence like the one presented below (15), the word *antivírus* (antivirus), that in the majority of the contexts would be a common noun, here refers to the name of a menu option, and, because of that, it was identified as a NE of the type CMD – despite the word not beginning with a capital letter (see the capitalization issue ahead).

(15) "Clique onde diz **antivírus**."
"Click where it says **antivirus**."

In example (16), the expression *Preto e Branco* (Black and White), that in another context could be a multiword expression, in this context again refers to the name of a menu option, and that is why it was also considered a NE of the type CMD.

(16) "Sim, faça a pesquisa depois vá até **Ferramentas de pesquisa** no separador **Cor** e escolha **Preto e Branco**"
"Yes, do the research and then go to **Research tools** in the tab **Color** and chose **Black and White**."

Another challenge arises from the fact that this annotation task was built on texts written by users and operators of an ICT support helpline, a corpus that is a set of spontaneous productions, without any kind of revision or adaptation. These texts are written in an informal style, since the goal is only to convey a message as fast and directly as possible. As such, it is common to find spelling errors, disrespect of capitalization rules, lack of punctuation or accentuation, among other problems that can be an obstacle to the annotation (as they are general obstacles to the comprehension of a text).

Regarding NEs in particular, the capitalization issue becomes particularly relevant. In Portuguese "proper names are conventionally spelled with an initial capital letter" (*idem, ibidem)*[4], a rule that is constantly overlooked in this corpus, not only by helpline users, but also by the operators who answer the questions (see example (15), taken from the part of the corpus that contains the answers). Thus, the NE identification cannot rely in the use (or not) of capital letters – the context, once again, becomes central to the annotator.

3.3 Results

In the NERC task, 9 477 NEs were found, which cover 16 002 (23%) of the 68 928 tokens in the already annotated portion of corpus. Given that, as has been mentioned before, the corpus is composed of question-answer pairs – meaning that the number of questions is the same as the number of answers –, we looked at the data and we created two groups: the set of all the questions on one hand, and the set of all the answers on the other hand. In Table 1 we can find the results.

	Total of tokens	Total of NEs	Tokens marked as NEs
Questions	31 161	3 253	4 904 (16%)
Answers	59 707	6 224	11 098 (19%)

Table 1: Constitution of the annotated corpus

Results show that in the set of answers the NEs number is higher as well as the number of tokens marked as NEs. Since there are 3 000 annotated question-answer pairs, and taking into

4 Our translation of "os nomes próprios são convencionalmente grafados com letra maiúscula inicial" (*idem, ibidem*: 1009).

account the number of NEs in each set, we can conclude that almost every questions and answers have at least one NE.

This number highlights the necessity of a specific treatment of NEs, so that this corpus can be used to develop natural language processing tools. As Figure 2 shows, the majority of NEs are tagged with the three labels added to the tagset (which correspond to the three semantic types more related to the ICT domain), lending support to our initial motivations for creating NE semantic types that are specific to the domain of the corpus.

We can conclude that the corpus specificity leads us to the necessity of an adapted and detailed annotation (i.e. leads us to a corpus-driven approach), because only then is it possible to perceive which NEs types can be found in it. The survey of the most common NEs categories in the corpus is an asset to its utilization and analysis, within the QTLeap Project and in the development of tools which support machine translation in particular and natural language processing in general.

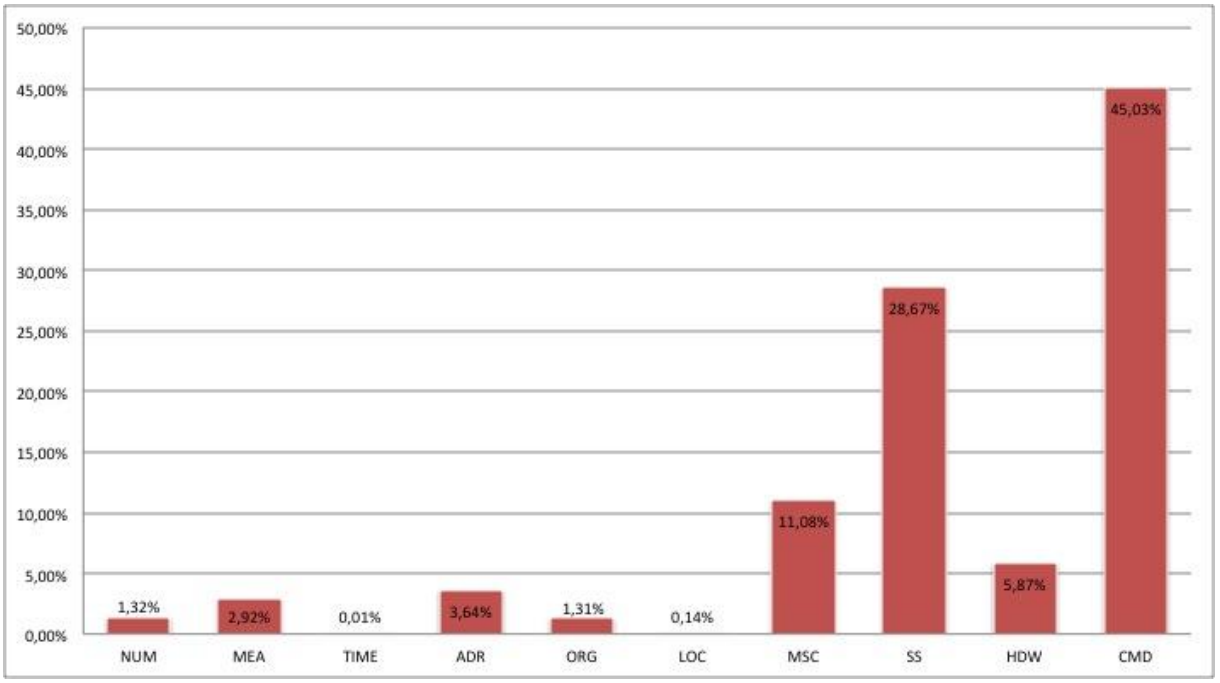

Figure 2: NEs distribution by type (second phase)

4. Named Entities Disambiguation task

4.1 Methodology

4.1.1 Annotation Task

After completing the first annotation task, NERC, which involved marking each NE and annotating it with its semantic type, a second task was carried out in which each NE is disambiguated by identifying the entity that it designates. This task is called Named Entity Disambiguation (NED) and consists of associating each NE to an entry in the Portuguese DBpedia (Lehmann *et al.*, 2015). DBpedia is a database obtained from the structured information in Wikipedia. Since each entry in DBpedia corresponds to a Wikipedia page and, since different entities have different pages, linking a NE to the DBpedia entry of its referent effectively disambiguates that NE. We use DBpedia version 3.9, generated from Wikipedia dumps in late March and early April of 2013.

To facilitate the task of the human annotator, before the manual disambiguation task, a pre-annotation script was used to automatically assign a tentative DBpedia entry to each NE being disambiguated. The annotators then had to verify whether the automatic disambiguation was correct and, if found to be wrong, choose the right DBpedia entry for the NE at stake. Throughout the task, the annotators had to take into account the NE semantic type assigned during the NERC task and the discourse context in which the entity occurs. Frequently, the same word can be associated to different entries from DBpedia, depending on the specific semantic type of the entity – for example, the word "Google" with the type ORG (organization) should be connected to the DBpedia page of the Google Company, while the same word with the type SS (software and services) should be connected to the DBpedia page about Google's search engine. For this NED task, the corpus was split into disjoint sets and each set was assigned to a single annotator. In future, we intend to repeat the annotations in each set, with a different annotator, and introduce the process of inter-annotator agreement and adjudication, so that the reliability of the resource can be quantified.

For the NED task, only the semantic types EVT, LOC, ORG, PER, WRK, MSC, SS and HDW were considered because, for the remaining types (NUM, MEA, TIME, ADR, CMD), we

hardly find an entry in DBpedia that can be considered to be the referent of those entities. From the 9 477 NEs found during the NERC task, 3 903 were disambiguated.

4.1.2 Annotation tool

The NED task was carried out by using the brat annotation tool (Stenetorp et al., 2012a). Like WebAnno, brat is also a web-based annotation tool that can be used through a browser, making its use very intuitive and simple. The brat tool was selected for this task due to its support for entity normalization (Stenetorp et al., 2012b), that is, the linking of any annotated item to an external resource – in our case, this allows linking any identified NE to its correspondent entry in DBpedia.

4.2 Challenges

Named Entity Disambiguation is a task that raises some problems. One of the main challenges is the linking of specific or abstract entities to a Portuguese DBpedia entry.

We have adopted an annotation approach according to which the annotator, when dealing with an ambiguous or unspecific NE, must link it to the DBpedia entry containing the most useful information for the denotation of the entity expressed by that NE. This was specially relevant in the cases of NEs that denote an entity with very specific characteristics, such as the version of a software product or the model of a piece of hardware, whose specific DBpedia entry often did not exist at all. An example of this is the NE "Acer Aspire p503m", denoting a specific computer model, which was disambiguated by linking it to the DBpedia entry corresponding to the family of models "Acer Aspire" computers, since the Portuguese DBpedia does not include any entry for the specific computer model.[5] The same approach was taken for the NE "Excel 2013", for example, which was disambiguated through its connection to the "Microsoft Excel" entry, presenting a more general version of this software (that is, its hypernym), instead of the specific 2013 version, which is not present in the Portuguese DBpedia.

5 If the DBpedia entry for the "Acer Aspire" family of models had not existed, the annotator would next try to link to the entry for "Acer" computers in general.

The lack of Portuguese DBpedia entries for certain NEs was one of the biggest challenges during the NED task. Note, for example, the illustrative cases of the NEs tagged as CMD, such as "F5" (the key) or "History" (a menu item), which do not have entries in that database. As explained above, we chose not to take into account the NEs belonging to certain semantic classes, like the ones with this tag CMD, since there are no possible DBpedia entries for them.

This being said, in order to highlight those cases of incomplete or problematic annotation, a set of annotator comments was established. These comments can be added to a NE to describe the type of problem that caused the impossibility of annotation, for a future re-annotation of this corpus to be facilitated. These comments include:

- "Not found - *page ID*", when the entity was not found in the Portuguese DBpedia but the annotator found it in the Portuguese Wikipedia. Note that this can happen because the version of DBpedia being used was generated in 2013, and many pages have been added to Wikipedia since. Adding the Wikipedia page ID to the comment of that NE will allow to automatically disambiguate it when, in the future, the version of DBpedia in use is upgraded.

- "Not found", for cases when no referent to the entity is found in DBpedia or Wikipedia.

- "Not found - EN", when there is no entry in the Portuguese DBpedia or Portuguese Wikipedia, but there is an entry in the English Wikipedia.

- "Typo", when the entity contains a spelling or segmentation error but its referent can still be recognized and therefore disambiguated.
- "Bad NERC", when the entity is wrongly classified or has a spelling mistake, and its referent cannot be identified by the annotator.

Another issue is contingent on the use of brat for this task. Searching the disambiguation database usually returns a very large number of DBpedia entries. For example, if the annotator wants to disambiguate the named entity "Google", brat collects, from DBpedia, all entries whose

title contains the word "google", such as "Google Earth", "Google Maps" and "Google", and it provides all of them as disambiguation options, in a list with no specific order. This way, the human annotator has access to a large number of possible DBpedia entries for the NE being disambiguated. However, since those entries are not sorted, either alphabetically or by the ID of the entry, the choice of the right entry becomes harder. To solve this problem, the annotator must sometimes search manually on the Wikipedia site, find the page for the NE referent, find the ID of that page and insert it manually into a field brat provides for the page ID, thus enabling the correspondence between that ID and the matching DBpedia entry. We find that it would be far more efficient for this task either to trim the number of entries returned by brat or at the very least sort them by ID, in order to facilitate easier search for particular named entities and easier association of a named entity to some DBpedia entry.

4.3 Results

Out of the 3 903 NEs chosen for disambiguation in the NED task, 3 345 (86%) were disambiguated. From the 558 NEs that were not possible to disambiguate, 282 were not found in DBpedia or in Wikipedia, and 197 were registered as existing in the Portuguese Wikipedia, but not in the Portuguese DBpedia database (and so they were not disambiguated too). Regarding these latter cases, we can say that their disambiguation is pending, but can be done automatically as soon as future releases DBpedia come to include these entries.

5. Final remarks

In this paper we presented the named entity annotation of QTLeap Corpus. The main goal was the creation of an annotated corpus with lexical and semantic information on named entities by marking, classifying and disambiguating named entities found in the text.

Firstly, we presented the manual named entities recognition and classification task (NERC), that consist in recognizing the named entities in the corpus and in tagging them with one of 13 semantic types (NUM; MEA; TIME; ADR; PER; ORG; LOC; EVT; WRK; MSC; CMD; SS; HDW). The majority of the named entities (79.57%) are classified with one of the three new tags that were created specifically for this corpus (CMD; SS; HDW).

Afterwards, we described the manual named entities disambiguation task (NED), in which we disambiguate entities by linking them to the corresponding entry in the Portuguese DBpedia. Approximately 86% of the named entities found in QTLeap corpus were disambiguated.

The main contribution of this work is a corpus of annotated and disambiguated named entities in a specific domain and genre, which can be used as a resource in the tasks of NERC (Named Entity Recognition and Classification), NER (Named Entity Recognition) or NED (Named Entity Disambiguation) of Portuguese. Given that the corpus is multilingual and aligned, this corpus can be the first step in the creation of a domain specific multilingual aligned lexicon, which is a valuable resource for the improvement of machine translation tools.

Acknowledgments

This work was partly funded by the Portuguese Foundation for Science and Technology through the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012), by the European Commission through project QTLeap (EC/FP7/610516) and by the European Commission through ASSET, Programa Portugal2020, co-promotion projects, Contract: 3279.

References

Babych, Bogdan, Anthony Hartley (2003) Improving Machine Translation Quality with Automatic Named Entity Recognition. In Proceedings of the 7th International EAMT Workshop on MT and Other Language Tools, pp. 1-8.

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar do Nascimento, Filipe Nunes, João Silva (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).

Del Gaudio, Rosa, Aljoscha Burchardt, Arle Lommel (2015) Evaluating a Machine Translation System in a Technical Support Scenario. In *Proceedings of the 1st Deep Machine Translation Workshop*, pp. 12-19.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer (2015) DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, Vol. 6 No. 2.

Mateus, Maria Helena Mira, Ana Maria Brito, Inês Duarte, Isabel Hub Faria, *et al.* (2003) *Gramática da Língua Portuguesa*. Caminho: Lisboa (5ªed).

Paiva Raposo, Eduardo Buzaglo & Maria Fernanda Bacelar do Nascimento (2013) Nomes próprios. In Raposo, Eduardo Paiva, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura, Amália Mendes (orgs) *Gramática do Português.* vol. I Lisboa: Fundação Calouste Gulbenkian.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii (2012) brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 102-107.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, Chris Biemann (2013) *WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations*. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (*ACL), pp. 1-6.