CORPUS NE

1. BASIC INFORMATION

1. *Corpus composition*

This corpus consists of a set of nearly 5,500 manually annotated questions to be used as training corpus in machine learning based NER systems and 500 annotated questions for testing. Named entities in these questions were identified and classified according to the categories: `Person`, `Location` and `Organization`. More details about the process of building these corpora can be consulted in [1].
The original corpus of 6000 questions in English can be found in http://cogcomp.cs.illinois.edu/Data/QA/QC/. Details about it are presented in [2].

2. *Representation of the corpora (flat files, database, markup)*

The corpus is a txt file.

3. *Character encoding*

The characters are encoded in UTF-8.

2. ADMINISTRATIVE INFORMATION

1. *Contact person*

Name:Luísa Coheur
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Assistant Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: luisa.coheur@inesc-id.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 Copyright statement and information on IPR

The resource is free.

3. TECHNICAL INFORMATION

1. *Directories and files*

The archive that will be uploaded on the MetaShare platform will contain one folder with two files with .txt extension (the train and the test file).

2. *Data structure of an entry*

This is not relevant as the corpus is provided as a text file, where each line contains a single question. In each question, all the named entities are identified: each named entity is between symbols "<" and ">".

3. *Corpora size (nmb. of tokens, MB occupied on disk)*

The corpus for training has 5452 questions, 55620 tokens and occupies about 315 KB. The corpus for testing has 500 questions, 3758 tokens and occupies about 25 KB.

4. CONTENT INFORMATION

1. *Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)*

This corpus is in monolingual.

2. *The natural language(s) of the corpus*

The language of the corpus is English.

4. 3 Domain(s)/register(s) of the corpus

The corpus has questions from different types: factoids, definitional, lists.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is annotated according with the following categories: `Person`, `Location` and `Organization`. An example can be seen in the following:

How did serfdom develop in and then leave <LOC>Russia</LOC> ?
What films featured the character <PER>Popeye Doyle</PER> ?
When did <ORG>CNN</ORG> begin broadcasting ?

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

Tags in use are:
a) <LOC> for Locations;
b) <PER> for Persons;
c) <ORG> for organization.

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not relevant

4.4.4 Attributes and their values (if annotated)

Not relevant

5. *Intended application of the corpus*

This corpus can be used to train and test Named Entity Recognition in questions. As questions are different from declarative sentences, they need an appropriate corpus for training, as showed in [1]. In addition, the corpus where we have identified the named entities is widely used by the machine learning community, because each of its questions is labeled based on one of the most widely known taxonomies for question classification: Li and Roth's two-layer taxonomy [2]. This taxonomy consists of a set of six coarse-grained categories and fifty fined-grained ones. This fact makes this corpus a very valuable resource for training and testing machine learning models in question classification, and, more generally, making it a very valuable resource for question answering. By identifying the named entities in that corpus, these can be used to improve the attained models, as named entities can also be used as features.

6. *Reliability of the annotations (automatically/manually assigned) – if any*

The annotations were automatically built and manually checked, being the annotation process described in [1].

5 RELEVANT REFERENCES AND OTHER INFORMATION

[1] Ana Cristina Mendes, Luísa Coheur, Paula Vaz Lobo. Named Entity Recognition in Questions: Towards a Golden Collection in LREC'10. May, 2010.

[2] Xin Li, Dan Roth, Learning Question Classifiers. COLING'02. August, 2002.