A multilingual Named Entity corpus for Arabic, English and French

Djamel Mostefa[1], Mariama Laïb[2], Stéphane Chaudiron[3], Khalid Choukri[1], Gaël de Chalendar[2]

[1]ELDA
55-57 rue Brillat Savarin Paris, France
{mostefa,choukri}@elda.org

[2]CEA-LIST
18, rue du panorama
BP6 92265 Fontenay aux Roses, France
meriama.laib@cea.fr

[3]Université de Lille3 – GERiiCO
Domaine univ. du Pont du Bois
BP 60149- 59653
Villeneuve d'Ascq cedex France
stephane.chaudiron@univ-lille3.fr

Abstract

This paper presents the semi-automatic annotation with Named Entities (NE) of a multilingual corpus. The languages are Arabic, English and French. The text corpus is made of comparable newswires from the Agence France Presse covering the period 2004-2006. Our method for producing the corpus is iterative. First the automatic tagging is produced by a state-of-the-art named entity tagger. Then the annotations are checked manually and corrected if necessary. The AFP corpus and annotation scheme are described. The paper presents also the statistics of the corpus and compare the annotation results for the three languages. The final corpus is made of 30, 000 tagged documents for the three languages, including 10,000 documents per language. The corpus is publicly available through ELRA's catalog of language resources.

Introduction and background

This article presents a multilingual corpus of Named Entities (NE) made of news wires from Agence France Presse (AFP) in three languages: Arabic, English and French. The project was supported by the French Ministry of Research.

The recognition and categorisation of person names, organisation names, location names, etc. is regarded as a fundamental process for a wide range of Natural Language Processing (NLP) tools and modules dealing with content analysis such as machine translation, information retrieval, information filtering, question answering, anonymization, etc.

There are few publicly available NE annotated corpora, especially for Arabic. Main resources come from evaluation campaigns on extraction of named entities.

Named entity extraction has been evaluated in various evaluation campaigns like MUC[1]-6/7 (Grishman and Sundheim, 1996) for English, MET[2] (Merchant et al, 1996) for Spanish, Japanese and Chinese, IREX[3] for Japanese (Sekine et Isahara, 1999), CoNLL[4] 2002-2003 for Spanish, Dutch, English and German (TjongKimSang, 2002), ESTER[5] for French (Galliano et al, 2006), HAREM[6] for Portuguese (Santos et al, 2006), ACE[7] 2000-2004 for Arabic, Chinese and English (Doddington et al, 2004). Only few of these corpora developed within evaluation campaigns are publicly available and none of them are multilingual. So our motivation was to develop and make publicly available a named entity corpus build on comparable documents for different languages. We used a corpus of newswires provided by Agence France Presse (AFP). The corpus covers a 3 years period (2004-2006) and 3 languages: Arabic, English and French.

The paper first presents the AFP corpus and gives an overview of the annotation guidelines that were used. We then focus on the annotated corpus and give some statistics of occurrences for the various classes of named entities and compare the results for three languages.

Named entity tagset

Usually named entities refer to proper names (proper nouns) and numerical expressions (dates, amounts, etc). Different sets have been defined and used from a 4 category set (person, location, organisation, miscellaneous) to the Sekine hierarchical proposal (Sekine et al 2002) with more than 100 subtypes. Before starting the annotation a survey of the different tagset used in recent projects has been conducted. We then came up with a tagset of 5 higher classes and 11 subclasses. The main categories are Person, Organization, Location, Date and

[1] Message Understanding Conference
[2] Multilingual Entity Task
[3] Information Retrieval and Extracion eXercise
[4] Conference on Natural Language Learning
[5] Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques
[6] Avaliação de Reconhecedores de Entidades Mencionadas
[7] Automatic Content Extraction

Numex. Organization and Date are divided into 2 subcategories while Location is divided into 5 categories.

Person names

Are annotated as Person entities, word or groups of words referring to real persons, fictional characters or religious deities. We also annotate as Person: nicknames, alias, titles and roles when they refer to persons and are not ambiguous.

Organization names

Are tagged as Organization entities, words referring to organizational structures in different domains such as sport teams, political parties, enterprises, etc. According to different contexts where the entity appears, and in order to resolve ambiguities, this class is divided into two subclasses: Organization.Org and Organization.GPE (Geo Political Entity). The first one annotates organization names used as such and the second one annotates location names when refer to organisations such as "*France* had decided not to go ahead with the flight after *Britain*".

Location names

This class annotates all locations and is divided into seven subclasses:

- Location.Geo: geographical location such as lakes, seas, mountains, etc.
- Location.Fac: buldings and facilities,
- Location.Line: roads and motorways,
- Location.Post.Addr: street addresses
- Location.Url.Addr: emails and web addresses,
- Location.Tel.Addr: phone numbers and faxes,
- Location.GPE: this subclass annotates location names when the context is referring to organizations.

DateTime tags

This class annotates expressions of time. This class is divided into two subclasses:

- dateTime.Time: hours, minutes, periodes, etc.
- dateTime.date: relative or absolute dates.

Numex tags

This class is used for numerical entities such as: distance, speed, weight, age, etc.

This tagset is compatible with the information containing in this kind of corpora and can be interesting to use in technology watch applications.

The AFP corpus

The corpus is provided by the Agence France Presse (AFP).

We selected 3 languages (Arabic, English and French) and a 3 years period (2004-2006) which represents a collection of about one and half millions newswires for around 10 GB and 424 millions words, from which 10,000 documents of each language have been selected to be used for the annotation with named entities. News articles are encoded in XML format and follow the News Markup Language (NewsML) specifications. NewsML is an XML standard designed to provide a media-independent, multilingual structural framework for multimedia news.

Statistics of the whole corpus are depicted in Table 1.

Wire	*# documents*	*# words*
Arabic	254 847	56 M
French	448 660	123 M
English	758 578	245 M
Total	1 462 085	424 M

Table 1 Statistics of the AFP corpus

In term of words, the Arabic corpus contains 56 millions of running words. The English corpus is made of 245 millions of words. The French corpus contains 123 million of words.

Annotation procedure

The documents are annotated in two steps. First, the named entities are recognized automatically in the documents using LIMA (LIC2M Multilingual Analyzer), a linguistic analyzer developed by the CEA LIST (Besançon et al., 2003). The named entities recognizer is a module of this analyzer which uses a set of rules and a list of special triggers (e.g., Miss, Mr Mme, President, lake, corporation, etc.) to identify named entities[8] and their types. These rules were adapted in order to consider the tagset of the categories in this project. Rules are activated when a trigger is found in the text. Then, the recognizer verifies right and left contexts in order to identify named entities when they exist. For example, in the sentence "أكدت وزارة الداخلية العراقية", the word "وزارة" is considered as a trigger. In this case, the rule that allows the recognizer to identify "وزارة الداخلية العراقية" as an Organization name stipulates that: if this trigger is preceded directly by a verb and followed by one or two specific adjectives, then the sequence of these two or tree words is tagged as an Organization entity. According to the right context, this sequence of words can be tagged as a Location entity if the first word is preceded by prepositions as "فِي " or "إلَى " and nouns as "موقع" or "بناء". Before this named entities recognition process, words are analyzed and tagged with their grammatical categories.

Then the annotations are checked and corrected manually. We recruited six annotators (two per language) for the manual correction. The annotators worked 4 hours per day since the task needs a high level of concentration. At the

[8] Originally, our system is able to recognize five types of named entities: Person, Organization, Location, Numex and Timex.

end of the project, the annotators could validate 50 documents per half-day which means that the ratio is around 12 documents per working hour.

Annotation tool

The annotation tool Annoqt has been developed by the CEA. It is a generic annotation tool and is very easy to use. It has been developed using the Qt library. As such, it is lightweight and multiplatform. It has been used under Windows and Linux platforms but it should be usable on any platform supported by Qt, like MacOS or mobile devices.

Thanks to the Qt library, Annoqt is natively multilingual. It particularly handles gracefully right-to-left languages and non-latin scripts. Thus, it is able to deal with the three languages of the corpus.

Using Annoqt, the user annotates segments of text with colors by selecting them with the mouse and then clicking on the desired color. Each color is associated to a Named Entity type (Person, Numex, Organization...) using a XML configuration file. A graphical configuration editor is offered in order to simplify the setup of the configuration.

Figure 2 shows an Arabic text being annotated. One can see on the left side the list of colored named entities types and the main annotation window on the right side. If the user clicks on an existing annotation, it selects it and he or she can then remove the annotation using the scissors tool or change its type by choosing a new one in the left panel list. You can see that Annoqt also supports overlapping entities (or even entities completely embedded inside other entities). This is shown by text segments decorated with colors gradients using the colors of all the entities types involved. When the user clicks with the left mouse button on such a segment, a contextual menu is displayed allowing him or her to select one of the entities.

Annoqt offers a few other features, like reusing the last used entity type on a new text segment or repeating the same annotation on all same strings of the text.

This is the first version of Annoqt. Future versions will be made easier to use in other projects by making pluggable the annotation format that is currently fixed. We also hope to be able to quickly release it under a Free Software license.

Corpus statistics

We selected the most recent 10,000 news wires per language from the whole AFP corpus for the annotation.

More precisely the English corpus is made of news from November 15 to November 29, 2006 while the French corpus contains news from October 22 to November 30, 2006 and the Arabic one range from October 1 to November 28, 2006.

At the beginning of the project, we planned to manually annotate the whole corpus composed of 30 000 documents. But the effort was much more important than expected and therefore we couldn't annotate the whole corpus within the project. In total, 5 278 documents were manually annotated, 1 177 for English, 1 785 for French and 2 316 for Arabic.

The 24 722 remaining documents are automatically annotated but couldn't be corrected manually.

Table 2 shows some statistics on the automatically annotated corpus and the manual one.

	Automatic annotation			Manual annotation		
	#doc	#word	#NE	#doc	#word	#NE
Ara	10k	2.4 M	190k	2,316	538k	106k
Eng	10k	3.4 M	261k	1,177	393k	43k
Fre	10k	2.9 M	228k	1,785	517k	78k

Table 2 Statistics on the automatic annotated corpus and the manual annotated corpus

We present here some statistics on the manually annotated part of the corpus.

We can see that the corpus is very rich in named entities. The English part contains 43,124 occurrences of named entities from 5,667 different entities. For French there are 78,442 occurrences from 10,377 named entities. For Arabic 106,434 occurrences of named entities were annotated. So the average number of named entities per document is 36 for English, 44 for French and 46 for Arabic.

For the three languages, the classification between categories is identical. The most frequent category is Location, then Organization, Person, Date and Numex.

Figure 1 gives an overview of the classification for the Arabic corpus. The repartition is not very surprising since the AFP corpus deals with news in general and therefore Geo Political Entities (GPE) are very present and are annotated either as Location.GPE or Ogranization.GPE.

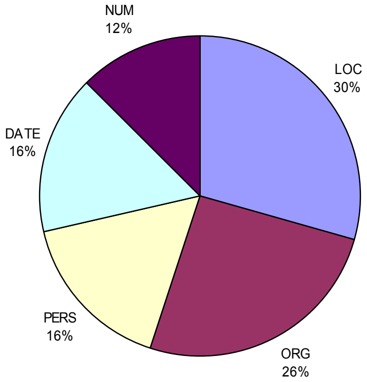


Figure 1 Named entities repartition for the Arabic news wires

Conclusion

In this paper we presented the corpus of named entities build on the AFP corpus. The database is made of 30 000 news wires in Arabic, English and French. The corpus was annotated semi automatically with the help of a state-of-the-art NE tagger. We provided detailed information on the annotation scheme, the content of the corpus and statistics of occurrences of NE tags.

In order to minimize the cost of the manual validation, we have to adapt the rules used in our system so that the quality of the automatic recognition is improved. This is what we plan to do in the next step of the project using the corpus already tagged and checked. We plan to tag 30 000 documents in each language.

The corpus is publicly available through ELRA's catalog of language resources (http://catalog.elra.info) for research or commercial use.

Acknowledgements

The project is funded by the French Ministry of Research.

Bibliographical References

Besançon R., de Chalendar G., Ferret O., Fluhr C., Mesnard O., and Naets H. (2003) The LIC2M's CLEF 2003 system. In Working Notes for the CLEF 2003 Workshop, August 2003.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S. and Weischedel R. (2004). The Automatic Content Extraction (ACE) program tasks, data, and evaluation. In Proceedings of the 4th international Conference on Language Resources and Evaluation (LREC 2004), Lisboa, Portugal.

Galliano S., Geoffrois E., Gravier G., Bonastre J.F., Mostefa D. and Choukri K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy.

Grishman, R and Sundheim, B, (1996). Message Understanding Conference-6: a brief history. In Proceedings of the 16th Conference on Computational linguistics. Morristown, NJ, USA.

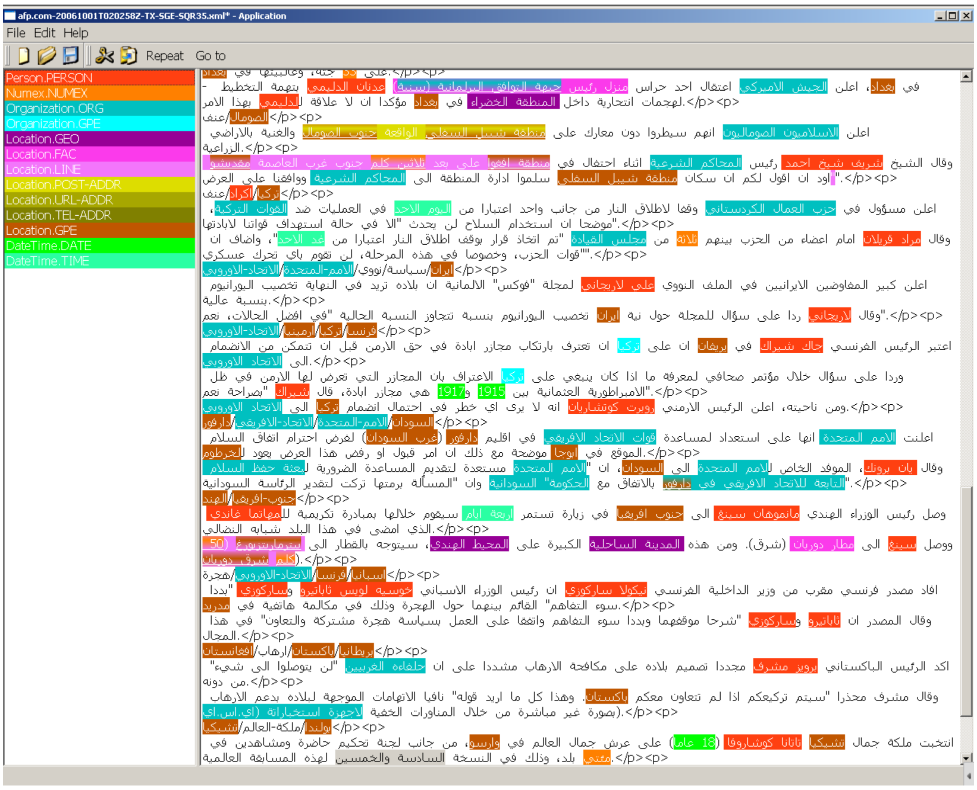
Merchant, R., Okurowski, M. and Chinchor, N. (1996). The multilingual entity task (MET) overview. In Proceedings of a workshop on held at Vienna, Virginia, pages 445–447, Morristown, NJ, USA. Association for Computational Linguistics.

Santos D., Seco,N., Cardoso N. and Vilela, R. (2006). HAREM : An Advanced NER Evaluation Contest for Portuguese. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy.

Sekine S. and Isahara H. (1999). IREX project overview. In Proceedings of the Information Retrieval and Extraction Exercise, Japan.

Sekine, S. Sudo, K. and Nobata, C. (2002) Extended Named Entity Hierarchy. In Proceedings of 3rd International Conference on Language Resources and Evaluation. Las Palmas, Spain.

TjongKimSang E. (2002). Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In Proceedings of the 6th conference on Natural language learning, Morristown, NJ, USA.



على 33 جثة، وغالبيتها في بغداد.</p><p>
في بغداد، اعلن الجيش الاميركي اعتقال احد حراس منزل رئيس جبهة التوافق البرلمانية (سنية) عدنان الدليمي بتهمة التخطيط -
لهجمات انتحارية داخل المنطقة الخضراء في بغداد مؤكدا ان لا علاقة للدليمي بهذا الامر.</p><p>
عنف/الصومال</p><p>
اعلن الاسلاميون الصوماليون انهم سيطروا دون معارك على منطقة شبيل السفلى الواقعة جنوب الصومال والغنية بالاراضي
الزراعية.</p><p>
وقال الشيخ شريف شيخ احمد رئيس المحاكم الشرعية اثناء احتفال في منطقة افغوا على بعد ثلاثين كلم جنوب غرب العاصمة مقديشو
"اود ان اقول لكم ان سكان منطقة شبيل السفلى سلموا ادارة المنطقة الى المحاكم الشرعية ووافقنا على العرض".</p><p>
عنف/اكراد/تركيا</p><p>
اعلن مسؤول في حزب العمال الكردستاني وقفا لاطلاق النار من جانب واحد اعتبارا من اليوم الاحد في العمليات ضد القوات التركية،
موضحا ان استخدام السلاح لن يحدث "الا في حالة استهداف قواتنا لابادتها".</p><p>
وقال مراد قريلان امام اعضاء من الحزب بينهم ثلاثة من مجلس القيادة "تم اتخاذ قرار بوقف اطلاق النار اعتبارا من غد الاحد"، واضاف ان
"قوات الحزب، وخصوصا في هذه المرحلة، لن تقوم باي تحرك عسكري"".</p><p>
الاتحاد-الاوروبي/الامم-المتحدة/سياسة/نووي/ايران</p><p>
اعلن كبير المفاوضين الايرانيين في الملف النووي علي لاريجاني لمجلة "فوكس" الالمانية ان بلاده تريد في النهاية تخصيب اليورانيوم
بنسبة عالية.</p><p>
وقال لاريجاني ردا على سؤال للمجلة حول نية ايران تخصيب اليورانيوم بنسبة تتجاوز النسبة الحالية "في افضل الحالات، نعم".</p><p>
الاتحاد-الاوروبي/ارمينيا/تركيا/فرنسا</p><p>
اعتبر الرئيس الفرنسي جاك شيراك في يريفان ان على تركيا ان تعترف بارتكاب مجازر ابادة في حق الارمن قبل ان تتمكن من الانضمام
الى الاتحاد الاوروبي.</p><p>
وردا على سؤال خلال مؤتمر صحافي لمعرفة ما اذا كان ينبغي على تركيا الاعتراف بان المجازر التي تعرض لها الارمن في ظل
الامبراطورية العثمانية بين 1915 و1917 هي مجازر ابادة، قال شيراك "بصراحة نعم".</p><p>
ومن ناحيته، اعلن الرئيس الارمني روبرت كوتشاريان انه لا يرى اي خطر في احتمال انضمام تركيا الى الاتحاد الاوروبي.</p><p>
دارفور/الاتحاد-الافريقي/الامم-المتحدة/السودان</p><p>
اعلنت الامم المتحدة انها على استعداد لمساعدة قوات الاتحاد الافريقي في اقليم دارفور (غرب السودان) لفرض احترام اتفاق السلام
الموقع في ابوجا موضحة مع ذلك ان امر قبول او رفض هذا العرض يعود للخرطوم.</p><p>
وقال يان برونك، الموفد الخاص للامم المتحدة الى السودان، ان "الامم المتحدة مستعدة لتقديم المساعدة الضرورية لبعثة حفظ السلام
التابعة للاتحاد الافريقي في دارفور بالاتفاق مع الحكومة" السودانية وان "المسألة برمتها تركت لتقدير الرئاسة السودانية".</p><p>
الهند/جنوب-افريقيا</p><p>
وصل رئيس الوزراء الهندي مانموهان سينغ الى جنوب افريقيا في زيارة تستمر اربعة ايام سيقوم خلالها بمبادرة تكريمية للمهاتما غاندي
الذي امضى في هذا البلد شبابه النضالي.</p><p>
ووصل سينغ الى مطار دوربان (شرق). ومن هذه المدينة الساحلية الكبيرة على المحيط الهندي، سيتوجه بالقطار الى بيترماريتزبورغ (50
كلم شرق دوربان).</p><p>
هجرة/الاتحاد-الاوروبي/فرنسا/اسبانيا</p><p>
افاد مصدر فرنسي مقرب من وزير الداخلية الفرنسي نيكولا ساركوزي ان رئيس الوزراء الاسباني خوسيه لويس ثاباتيرو وساركوزي "بددا
سوء التفاهم" القائم بينهما حول الهجرة وذلك في مكالمة هاتفية في مدريد.</p><p>
وقال المصدر ان ثاباتيرو وساركوزي "شرحا موقفهما وبددا سوء التفاهم واتفقا على العمل بسياسة هجرة مشتركة والتعاون" في هذا
المجال.</p><p>
افغانستان/ارهاب/باكستان/بريطانيا</p><p>
اكد الرئيس الباكستاني برويز مشرف مجددا تصميم بلاده على مكافحة الارهاب مشددا على ان حلفاءه الغربيين "لن يتوصلوا الى شيء"
من دونه.</p><p>
وقال مشرف محذرا "سيتم تركيعكم اذا لم تتعاون معكم باكستان. وهذا كل ما اريد قوله" نافيا الاتهامات الموجهة لبلاده بدعم الارهاب
لاجهزة استخباراتة (اي.اس.اي) بصورة غير مباشرة من خلال المناورات الخفية.</p><p>
تشيكيا/ملكة-العالم/بولندا</p><p>
انتخبت ملكة جمال تشيكيا تاتانا كوشاروفا (18 عاما) على عرش جمال العالم في وارسو، من جانب لجنة تحكيم حاضرة ومشاهدين في
مئتي بلد، وذلك في النسخة السادسة والخمسين لهذه المسابقة العالمية.</p><p>

Figure 2 Screenshot of the annotation tool Annoqt