Aleda, a free large-scale entity database for French

Benoît Sagot[1], **Rosa Stern**[1,2]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France
2. Agence France-Presse – Medialab, 2 place de la Bourse, 75002 Paris, France
benoit.sagot@inria.fr, rosa.stern@afp.com

Abstract

Named entity recognition, which focuses on the identification of the span and type of named entity mentions in texts, has drawn the attention of the NLP community for a long time. However, many real-life applications need to know which real entity each mention refers to. For such a purpose, often refered to as entity resolution and linking, an inventory of entities is required in order to constitute a reference. In this paper, we describe how we extracted such a resource for French from freely available resources (the French Wikipedia and the GeoNames database). We describe the results of an instrinsic evaluation of the resulting entity database, named Aleda, as well as those of a task-based evaluation in the context of a named entity detection system. We also compare it with the NLGbAse database (Charton and Torres-Moreno, 2010), a resource with similar objectives.

Keywords: Named Entities, Entity Database, Named Entity Linking

1. Introduction

Identifying named entities (NE) is a widely studied issue in Natural Language Processing, because NEs are crucial targets in information extraction or retrieval tasks, but also for preparing further NLP tasks (e.g., parsing). Therefore a vast amount of work has been published that is dedicated to NE*recognition*, i.e., the task of identification of NE *mentions* (spans of text denoting a NE), and sometimes *types*. However, real-life applications need not only identify NE mentions, but also know which real entity they refer to; this issue is addressed in tasks such as knowledge base population with entity resolution and linking. For such a purpose, an inventory of entities is required prior to those tasks in order to constitute a reference.

In this paper, we introduce such an entity database for French, Aleda, which is large-scale, freely available, and automatically extracted from free resources. After a short description of the related work in Section 2, with an emphasis on the entity database NLGbAse (Charton and Torres-Moreno, 2010), we describe in Section 3 how we extracted Aleda from the French Wikipedia and the GeoNames database. Finally, we evaluate Aleda both instrinsically and in a task-oriented setting (Section 4).

2. Related work and contribution

Aleda is related to a number of works through its extensive use of Wikipedia as a base resource for entities. In particular, the tasks of entity disambiguation, linking and knowledge population have often relied on encyclopedic information about entities provided by Wikipedia. Following (Bunescu and Pasca, 2006), a lot of research has been conducted in order to efficiently exploit both the content and the structure of Wikipedia (Cucerzan, 2007; Giuliano and Gliozzo, 2008; Dredze et al., 2010; Zheng et al., 2010). This is usually achieved along the lines of the following two-step process:

1. selecting the Wikipedia articles corresponding to entities, often by examinating the form of their titles, and
2. associating with the selected entities a set of attributes extracted from the Wikipedia structure: entity variants from text anchors, redirections and disambiguation pages, topic relation from article categories, lexical associations using the article texts and co-occuring entities.

The way Wikipedia is used for building Aleda (see Section 3.2) differs from this standard procedure, mostly (i) in the way entities are selected (implicitly through typing rather than by title form examination) and (ii) by the reduced amount of knowledge finally associated with entities. This is due to the fact that Aleda is not designed for a particular task but rather aims at being a stand-alone resource whose structure can be used by different applications with minimum adaptation. Aleda, which can be seen as a Wikipedia and GeoNames extraction rather than exploration, can thus be used directly in any application, avoiding potentially impeding computing costs often encountered when exploiting resources such as Wikipedia on-the-fly.

Another stand-alone base extracted from the French Wikipedia exists, namely NLGbAse (Charton and Torres-Moreno, 2010). The extraction process used for creating NLGbAse is comparable to our own approach for building Aleda. NLGbAse is not limited to NEs but provides a conceptual structure by grouping entities in main types (LOC, PERS, etc.) and subtypes (PERS.HUM for humans, LOC.ADMI for administrative locations, etc.). It contains 1,085,812 entries, whose distribution accross main types is shown in Table 1. Note that mentions in NLGbAse are titles of Wikipedia articles; for example, *Paul Ier (pape)* is a mention for the entity whose English standard name is *Pope Paul I*, although "*(pape)*" will probably never been found after *Paul Ier* in a real text. Another issue with NLGbAse is the fact that almost one fifth of all entries have an unknown type (see Table 1). Such entries sometimes correspond to entities which should have been typed (e.g., *musée de l'érotisme* which should have been typed as an organization), often to terms that are not NEs (e.g., notions such as

	#entities	#mentions
Person	326,938	488,328
Location	276,872	450,284
Product	142,428	236,611
Organization	119,487	210,898
Time	9,632	12,064
Function	1,954	3,942
Unknown	208,501	343,369
total	*1,085,812*	*1,864,754*

Table 1: Quantitative information about the NLGbAse database (Charton and Torres-Moreno, 2010).

dommage collatéral ’collateral damage’ or *cyclotron* ’cyclotron’, names of species such as *oreillard* ’plecotus’), but also to other types of pages (e.g., *liste swadesh de l’arabe* ’Swadesh list for Arabic’).

3. Building Aleda

Since the preliminary version of Aleda, briefly described in (Stern and Sagot, 2010), the extraction process from the French Wikipedia and GeoNames has been strongly refined and the resulting database has been evaluated and used in various settings. Each entity in Aleda is associated with information such as a weight or a subtype, as well as with a set of variants (mentions), themselves associated with additional information (e.g., the first/middle/last name structure for most person names).

3.1. Extracting locations from GeoNames

For location names, the extraction process from GeoNames is fairly straightforward. We first extract from the huge GeoNames database all entities associated with a subtype considered relevant for domain-generic corpus processing in French (countries, cities and so on). However, this filtering still outputs too many entities. Therefore, we designed heuristics for selecting entities considered as relevant for processing such corpora (e.g., we retained all locations in France as well as all non-French locations for which GeoNames provide a number of inhabitant greater then 200; all mentions associated with a non-empty language tag other than French were discarded, as well as all mentions using non-latin characters). Each selected entity is extracted with its GeoNames id and URI, its standardized name and its geographical coordinates, as well as a weight computed from the number of its inhabitants.

3.2. Extracting persons and organizations from Wikipedia

For other types of entities, we designed an architecture for exploiting the French Wikipedia, following previous work (Balasuriya et al., 2009; Charton and Torres-Moreno, 2010). In order to select all relevant articles corresponding to entities, we proceed in three steps. All *infoboxes*[1] included in articles are first extracted. Only a small proportion of articles include an infobox, but most infobox templates used in articles corresponding to entities can be easily associated with a unique entity type. A type such as Person, Location or Organization has thus been manually assigned to the most frequent infobox templates (136 infoboxes). This allowed us to assign a type to 202,571 articles, i.e., relevant entities. In a second step, we look for wikipedia *categories* associated with these already typed articles. For each of these categories occuring at least four times, we count the number of times they appear in articles of each type. If a category appears mostly within articles of one type, this type is associated with the category. This happened to 20,328 categories (e.g., *Naissance en 1978* appears in 2,777 articles, 1,777 of which have an infobox, which in all cases have been assigned the type “Person”; therefore, this category is associated with the type “Person”). Entities are finally extracted as follows: for a given article, we extract the type of each of its categories that has been assigned one; the most frequent type is assigned to the article; if no type can be assigned this way, we look if the article contains an infobox; if it is the case and if this infobox was manually typed, the corresponding type is assigned to the article; if it is not the case, this article will not correspond to an entity in Aleda.

	#entities	#mentions
Person	272,507	966,793
Location	515,242	522,728
Organization	32,963	84,949
Company	11,740	23,356
Product	4,675	8,060
Work	34,960	67,316
total	*872,087*	*1,673,202*

Table 2: Quantitative information about Aleda

For each article which was selected as an entity, we extract the entity URI, a standardized name (the title of the article) and a weight computed from the number of lines of the wikipedia article, which can be seen as an indication of the entity’s popularity. Entity variants are extracted using Wikipedia redirection links, as well as disambiguation pages: their titles are considered as variants of entities corresponding to articles they are referring to as alternatives.[2] For person names, additional variants are automatically generated, by identifying the first, middle and last names (when relevant), based on the title of the article and its first line (e.g., the variants *J.W. Smith* and *John W. Smith* can be added to the entity named *John William Smith*).

In its last version, the Aleda entity database contains 872,087 entities associated with 1,673,202 mentions. The breakdown by major entity types is given in Table 2. A few examples of entries, together with their counterpart in NLGbAse (when it exists), is given in Table 3.

4. Evaluation and contextual use of Aleda

An entity base like Aleda can be evaluated with regard to both its intrinsic quality as a collection of entities and its usefullness when used for a particular task.

[1] An infobox is a fixed-format table added to articles to consistently present a summary of some unifying aspect that share sets of articles.

[2] This process requires to filter out other article titles the disambiguation page is pointing to.

4.1. Intrinsic evaluation

The quality of Aleda can be evaluated by (i) mesuring the rate of variants which are correctly linked to a particular entity, i.e., which are relevant variants of the considered entity (precision) and (ii) its coverage (recall), for which a reference is needed.

We evaluated the precision (i) by randomly selecting 100 variants from Aleda and checking whether the variant was relevant as such and whether the association with one or several entities was correct. Among those 100 cases, one was not a relevant entity variant and two were on the verge of irrelevance (but established redirections in Wikipedia, e.g., *Martino* for *Martino Longhi l'Ancien*); all 97 others were correct variants, with a correct link to the associated entity or entities.

For recall (ii), we use as a reference a manually annotated corpus described in (Stern and Sagot, 2010) and considered three different cases. This corpus contains 709 distinct entity mentions, corresponding to 610 distinct entities, 518 having a Wikipedia or GeoNames page.[3] Among them, 437 match an entity in Aleda (over 508 distinct mentions). Among the 709 distinct mentions in this corpus, 497 are in Aleda. Among the 617 distinct pairs of entity/mention where the entity has a Wikipedia or GeoNames page, 462 exist in Aleda. The coverage in terms of entities is thus evaluated at 84.4%, 70% in terms of mentions and 74.9% in terms of entity/mention pairs.

4.2. Task-based evaluation: using Aleda in a named-entity recognition system

Aleda has been used in the context of a knowledge base construction at the French press agency (AFP): for the purposes of content enrichment with metadata and integration of the AFP production in the Linked Open Data network,[4] a reference base of the most relevant entities for the AFP (e.g., person, location, organization names) is needed. In order to initiate the population of this KB, a large corpus of AFP news has been tagged with a NER system which use Aleda mentions as a lexicon for its detection rules.[5] Recognized entity mentions were then linked to matching entities in Aleda; finally, those links where manually validated (or simply rejected when the NER system provided a false mention) and resolved in the cases of ambiguities (several possible entities for one mention), where the information provided by Wikipedia and Geonames URIs could help the process of disambiguation. Thanks to the large-coverage of Aleda and its compliance with the Linked Data requirements,[6] this process provided a set of 5,350 domain related entities. A second iteration using a more recent corpus and the same method enabled to enrich this initial population with a set of 2,300 new entities, which illustrate the quality of Aleda's coverage. Although not fully automated, it allowed for a fast and efficient initialization of the KB with relevant and structured data. Future population will follow the same principles but will include an automatic entity linking module. This module will also use Aleda as an entity resource and thus link entity occurrences in news wires to Aleda; new entities can then be suggested as candidates to be integrated into the AFP KB, still using Aleda information related to Linked Data.

5. Conclusion and perspectives

We have sketched how we have extracted Aleda, a large-scale entity database for French from freely available sources, namely the French Wikipedia and the GeoNames database. We have described and illustrated Aleda and compared it briefely to NLGbAse (Charton and Torres-Moreno, 2010), another such database.

Apart from improving the accuracy of the extraction results, in particular concerning the "definition" associated with each entry, the perspectives of this work are three-fold. First, we aim at increasing the granularity of the structured information embedded in Aleda, for example by introducing sub-types for entities other than location names. Second, we intend to put together information about locations coming from Wikipedia and from GeoNames. Finally, we have already started building Aleda for languages other than French.

Acknowledgements

This work was funded by the French ANR national grant EDyLex (ANR-09-CORD-008).

6. References

D. Balasuriya, N. Ringland, J. Nothman, T. Murphy, and J. R. Curran. 2009. Named entity recognition in wikipedia. In *People's Web '09: Proceedings of the 2009 Workshop on The People's Web Meets NLP*, pages 10–18, Suntec, Singapour.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, pages 9–16.

E. Charton and J.M. Torres-Moreno. 2010. Nlgbase: a free linguistic resource for natural language processing systems. In *Proceedings of LREC 2010*, Valletta, Malta.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL 2007)*, pages 708–716.

M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics (ACL 2010)*, pages 277–285.

C. Giuliano and A. Gliozzo. 2008. Instance-based ontology population exploiting named-entity substitution. In *Proceedings of the 22nd International Conference on Computational Linguistics (ACL 2008)*, pages 265–272.

[3] Only those entities where considered for the evaluation of entity coverage.

[4] `http://linkeddata.org/`

[5] The NP system is briefly described in (Stern and Sagot, 2010).

[6] Indeed, every Aleda entity is associated with a Wikipedia or GeoNames URI, which in turn links with other Linked Data knowledge such as DBPedia and the New York Times LOD (`http://data.nytimes.com`).

Aleda			NLGbAse		
id and name	information	mentions	id	type	mentions
1000000000001054 Émile Benveniste	type=PERSON gender=m weight=15 def=Émile Benveniste est un linguiste français [. . .] URI=`http://fr.wikipedia.org/wiki/Émile Benveniste`	Benveniste E. Benveniste Emile Benveniste É. Benveniste Émile Benveniste	15784093	PERS.HUM	Émile Benveniste
1000000000685469 Enrique Iglesias	type=PERSON gender=m weight=21 def=Enrique Iglesias, de son nom complet Enrique Miguel [. . .] URI=`http://fr.wikipedia.org/wiki/Enrique Iglesias`	E. Iglesias Enrique Iglesias Iglesias	*not in NLGbAse*		
2000000000745044 Istanbul	type=LOCATION sub_type=City geonames_type=P.PPLA weight=11174257 country=TR coordinates=28.9,41.0 zoom_level=9.32 URI=`http://www.geonames.org/745044`	Istanbul Byzance	298152143	LOC.ADMI	Istanbul Istanboul Istanboul Islam-bol İstanbul Stambouliote Istambul İstanbuliote Istamboul
2000000003017382 Republic of France	type=LOCATION sub_type=Country geonames_type=A.PCLI weight=64768389 country=FR coordinates=2,46 zoom_level=20 URI=`http://www.geonames.org/3017382`	France Republic of France Republique Française	16391564	loc.admi	France Française La France France. La france France (pays) Françaises FRANCE
1000000003065020 Parti radical de gauche	type=ORGANIZATION weight=37 def=Le Parti radical de gauche (PRG) est un parti politique français[. . .] URI=`http://fr.wikipedia.org/wiki/Parti radical de gauche`	Parti radical de gauche Parti Radical de Gauche Parti Radical de gauche Parti radical de Gauche PRG Mouvement des Radicaux de Gauche Radicaux de gauche RG	5079845048	ORG.POL	Parti radical de gauche Parti Radical de gauche Parti radical de Gauche Radicaux de gauche Parti Radical de Gauche Mouvement des Radicaux de Gauche (parti)
1000000000025542 Conseil supérieur de l'audiovisuel	type=ORGANIZATION weight=38 def=Le Conseil supérieur de l'audiovisuel (CSA) est l'autorité de [. . .] URI=`http://fr.wikipedia.org/wiki/Conseil supérieur de l'audiovisuel (France)`	Conseil supérieur de l'audiovisuel CSA	241112520	org.gsp	Conseil supérieur de l'audiovisuel (France)
1000000000788982 Association professionnelle des magistrats	type=ORGANIZATION weight=5 def=L' Association professionnelle des magistrats ou APM est [. . .] URI=`http://fr.wikipedia.org/wiki/Association professionnelle des magistrats`	Association professionnelle des magistrats Association Professionnelle Des Magistrats APM	2415738661	PERS	association professionnelle des magistrats APM (homonym)
1000000003658728 Bank Indonesia	type=ORGANIZATION weight=1 def=La banque d'Indonésie est la banque centrale indonésienne URI=`http://fr.wikipedia.org/wiki/Bank Indonesia`	Bank Indonesia	*not in NLGbAse*		

Table 3: Examples of Aleda entries, compared with their counterpart in NLGbAse (Charton and Torres-Moreno, 2010). Note that Aleda mentions for person names are structured (first/middle/last name or nickname) although this is not shown in this table. NLGbAse DBpedia-based URIs are not shown either

R. Stern and B. Sagot. 2010. Resources for Named Entity Recognition and Resolution in News Wires. In *Entity 2010 Workshop at LREC 2010*, Valletta, Malta.

Z. Zheng, F. Li, M. Huang, and X. Zhu. 2010. Learning to link entities with knowledge base. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2010)*, pages 483–491.