Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data

Rosa Stern, Benoît Sagot

Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data

Rosa Stern
Univ. Paris 7, Sorbonne Paris Cité, France
INRIA-Alpage, Paris, France
AFP-Medialab, Paris, France
rosa.stern@afp.com

Benoît Sagot
INRIA-Alpage, Paris, France
benoit.sagot@inria.fr

Abstract

We present a practical use case of knowledge base (KB) population at the French news agency AFP. The target KB instances are entities relevant for news production and content enrichment. In order to acquire uniquely identified entities over news wires, i.e. textual data, and integrate the resulting KB in the Linked Data framework, a series of data models need to be aligned: Web data resources are harvested for creating a wide coverage entity database, which is in turn used to link entities to their mentions in French news wires. Finally, the extracted entities are selected for instantiation in the target KB. We describe our methodology along with the resources created and used for the target KB population.

1 An Entity Extraction Methodology for Knowledge Base Population

Current research conducted at the French news agency AFP focuses on the acquisition and storage of knowledge, particularly entities, relevant for the news production and usable as metadata for content enrichment. This objective sets off the need for a dedicated knowledge base (KB) relying on a light-weight ontology of entities mappable to the Linked Data framework. Identification of entities such as persons, organizations and geopolitical entities (GPEs)[1] in unstructured textual data, news wires in French in our case, underlie the construction and enrichment of such a KB.

This specific need is met by the KB *population task*, now well defined within the annual TAC dedicated track (Ji et al., 2011). KB population indeed relies on the ability to link entity mentions in textual data to a KB entry (*entity linking subtask*, henceforth EL),[2] which follows pioneer work in entity disambiguation (Bunescu and Pasca, 2006; Cucerzan, 2007). In a similar way to systems described in Dredze et al. (2010) and Ji & Grishman (2011), we conduct EL over AFP news wires in order to obtain relevant entities meant to populate the target KB, adapting these techniques to French data. This linking process is based on Web data extraction for both coverage and the purpose of Linked Data integration, which has become a widely explored trend in news management and publishing projects, such as the ones conducted by the BBC (Kobilarov et al., 2009) or the New York Times (NYT).

Compared to other KB population settings, this knowledge acquisition process is done throughout a sequence of resources and extraction steps rather than in a cyclic way. Instead of considering one KB as both an entity resource and the target of the population task, the target KB (AFP *Metadata Ontology*, henceforth AMO) is viewed as initially empty and progressively augmented with entity instances. This is because the use intended for AMO does not rely on exhaustivity, but on a relevant set of entities mentioned in the daily news production. This set is not fixed *a priori* and must be regularly updated in order to maintain a reflection of entities' emergence in the news domain. For instance, not all cities in the world

[1] These entity types are the usual focus of Information Extraction systems and are defined among others by the ACE entity recognition task (Doddington et al., 2004).

[2] The consequent *slot filling subtask* associates these entries to attributes and relations to other entities.

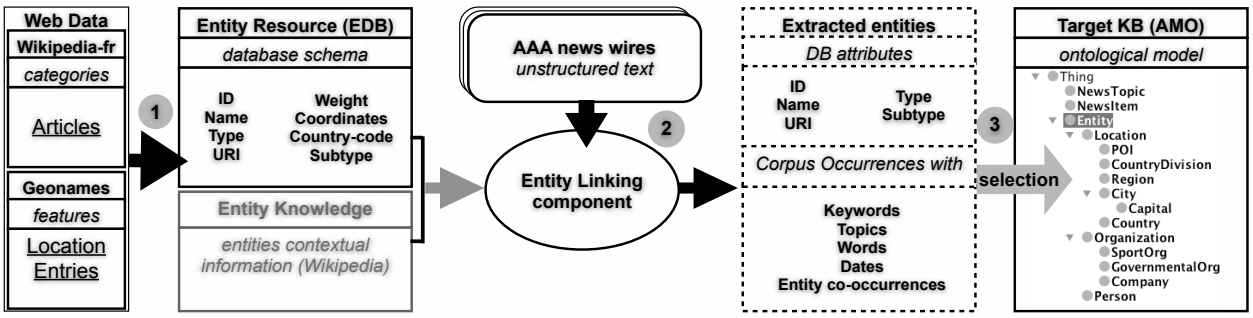

Figure 1: Overview of entity extraction and KB population process

need be instanciated in AMO, but the city *Fukushima* should become an entry as soon as its part in major events is stated in news wires. The relevant entity set can then be matched to new entries and updates in parallel documentation resources maintained at the AFP. The KB population process is broken down into several layers of data extraction and alignment, as sketched in Figure 1:

Step 1 The models of GeoNames and Wikipedia are mapped to a unified entity model; extraction from these Web datasets based on this mapping result in an entity database named Aleda, whose schema is presented in section 2, along with its alignement with Wikipedia and GeoNames data models.

Step 2 The Aleda database obtained in Step 1 provides entities for the linking step, where entity mentions in news wires are aligned with entries from Aleda. This process, along with the use of a joint resource for entity knowledge access, is described in section 3. EL in our particular task also targets the identification of new entities (i.e. absent from Aleda) and attempts to deal with the problem of possible named entity recognition errors in queries made to the system.

Step 3 The resulting entities aligned in Aleda must be anchored in the target KB, *via* instantiation of the adequate ontological class. Contextual information gathered during the entity extraction and linking process can be used at this point (section 4).

2 Entity Extraction from Web Data

Step 1 in our architecture is based on two Web datasets: the geographical database GeoNames brings together millions of location identifiers with associated information; Wikipedia is a wide coverage resource for entities, with continuous updates and article creations for entities involved in current events. The creation of a large-scale and unified entity resource is achieved by defining a database schema dedicated to entity representation, and by aligning both Wikipedia's and GeoNames' model with it. The schema considers *Person, Organization* and *GPE* as types of entries. The building of Aleda therefore relies on the identification of Wikipedia's and GeoNames' entries to which one of these types can be assigned.

Wikipedia Exploiting Wikipedia as a large-scale entity resource has been the focus of numerous efforts, such as (Balasuriya et al., 2009; Charton and Torres-Moreno, 2010). Each Wikipedia article, referring to an entity, concept or notion, is referenced under a number of *categories*, often fine-grained and specific. A hierarchical or ontological organization of these categories can be inferred (Syed et al., 2008) but a given article is not anchored in a generic conceptual class such as an entity type in a straightforward fashion. The present model alignment thus consists in a mapping from Wikipedia categories to one of the target entity types. Each article mapped to one of these types leads to adding a corresponding entity in Aleda. The selection and typing process of articles as entities makes use of heuristics based on articles' categories and *infoboxes* and is achieved as follows.

Around 100 of the most frequent infobox templates are retrieved from the articles' structure and manually associated with an entity type (e.g., the *Politician* template is typed as Person). All articles associated with one of these templates (23% of the French Wikipedia) are assigned the same type. The categories associated with these typed articles are then considered: a category appearing mostly in articles of a given type is associated with it. 20,328 categories are thus typed (e.g., all articles with an

infobox and tagged by a category such as *Born in ** have been assigned the type *Person*, which is therefore associated to this category). An article is selected as referring to an entity by assigning the type with maximal category association to it. If no type can be assigned (when no categories were associated with any type), an article's infobox whose template has been typed can provide the information. If the article has no such infobox, no entity is derived from it. Aleda's schema is filled with attribute values extracted from each selected article: the columns URI, name, weight are mapped to the article's URL, the normalized name inferred from the article's title and the article's length,[3] respectively. A joint table furthermore groups possible *variants* or *labels* for each entity; these are inferred from Wikipedia's redirection and disambiguation pages, which provide aliases for article titles. For person names, additional variants are automatically generated, by identifying the first, possible middle and last names.

GeoNames The model alignment from GeoNames is fairly straightforward since all its entries are locations. However this database is huge and present some noisy aspects, which we aim at avoiding by limiting the extraction to entities considered relevant for news content enrichment in French. Only GPEs such as countries, cities or country divisions were selected, based on the GeoNames *feature* provided. Heuristics are then designed for further selection: all GPEs in France are retained, as well as all non-French GPEs with more than 200 inhabitants. Aleda's schema is filled with values provided by GeoNames for each selected GPE: the columns URI, name, weight, subtype, country-code and coordinates are mapped to the GeoNames' entity URL, *name*, number of inhabitants, *feature* (such as P.PPL for populated place), ISO country code and coordinates, respectively. The joint variants table is filled with GeoNames' labels indicated as French or without language indication.

A unified entity resource of persons, organizations and GPEs[4] is obtained, with 832,452 uniquely identified entities (33%, 62% and 5% of type *Person, GPE* and *Organization*, respectively), associated with 1,673,202 variants.[5]

3 Text-Entity Alignment: Entity Linking

3.1 Methodology for Entity Linking

The knowledge acquisition step, crucial for the target KB population, consists in entity extraction from a corpus of AFP news wires. This is done by aligning detected entity mentions in news wires with entries the Aleda database introduced in 2. This linking component is based on the learning of a similarity model between entity attributes, including contextual ones, and a mention in a given document. It is challenging particularly because of name variation and ambiguity, which can prevent an accurate entity identification. The named entity recognition (NER) module first detects entity mentions, along with a possible type, which become queries for the candidates selection among the database entries. The entry with the highest similarity w.r.t. the mention and the document is chosen. This similarity is computed in a way comparable to a number of linking systems (described and evaluated in Dredze et al. (2010) and Ji & Grishman (2011)) and using the following features:

Surface similarity The candidate's name and labels, available in the entity database, are compared to the mention (exact, partial, null string match).

Candidate attributes The other candidate's attributes available in Aleda, such as its weight (or popularity) and its country-code, can be indicators in cases of ambiguity.

Contextual similarity Entity contextual knowledge (associated to the entity resource for the linking component in fig. 1) is made available from the corpus of Wikipedia articles. In each article, entity mentions (identified by hyperlinks) are surrounded with information such as the article's categories (i), the most salient words of the article's text (ii) and the co-occurring entities in the article (iii). As news items are indexed *via* keywords and topics from AFP's controlled vocabulary lists, (i) are mapped to the document keywords and topics; (ii) and (iii) are

[3] This value is seen as a weight in the extend that an article size can indicate the popularity of an entity relatively to others, particularly in cases of homonymy.

[4] GPEs extracted from both GeoNames and Wikipedia are listed and associated by an *owl:sameAs* relation.

[5] A comparison with the NLGbAse resource (Charton and Torres-Moreno, 2010), which has similar objectives, can be found in (Sagot and Stern, 2012). Aleda is freely available at `gforge.inria.fr/frs/download.php/30598/`

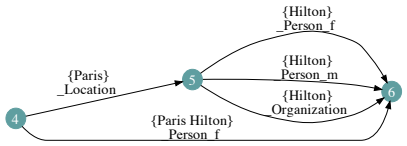

Figure 2: SXPipe/NP output for the segment *Paris Hilton*

compared to the document's salient words and entity mentions, respectively.

The candidate maximizing the similarity is selected as the mention linking. However, this selection should not occur for every query, since (i) a number of entities mentioned in news wires don't have a corresponding Aleda entry and (ii) the automatic NER can return a number of *false mentions*. In the case of (i), the *out-of-base* entity (NIL) should be output as a linking result; in (ii), the mention should be discarded and indicate a *not-an-entity* reading (NAE) rather than give rise to a false link; for the latter eventuality, features indicating ambiguities with the common lexicon are added to the feature set. These cases are part of the training examples and are taken into account in the prediction by including the NIL and NAE candidates in the ranking process.

3.2 Experiments and Evaluation

Experiments and evaluation are conducted over a manually annotated corpus of 96 news items, dated May-June 2009, where each entity mention's boundaries and Aleda ID, if relevant, are indicated. Mentions referring to entities absent from Aleda are identified by a normalized name. The corpus includes 1,476 entity mentions, referring to 610 distinct entities among which 28% are absent from Aleda. The corpus is also annotated with the NER system SXPipe/NP[6] to add examples of false matches to the linking model. More precisely, the NER is applied without complete disambiguation: a number of typing and segmentation alternatives are not resolved at this level, but rather passed to the linking module. The underlying idea is to leave final entity recognition decisions to a level where knowledge and semantic information about entities are available. Figure 2 illustrates this ambiguity preservation, where SXPipe/NP builds possible readings of a text segment in terms of entity mentions.

In each reading built by the NER module, each mention (gold or not) is associated with a set of candidates: all Aleda entries for which the mention is a possible variant, as well as NIL and NAE instances. All mention/candidate pairs are assigned a class, positive for correct links and negative for wrong ones, and form the training set for the linking model. A pair consisting in a false mention and the NAE candidate is for instance labeled as positive. The training examples are fed to a maximum entropy classifier[7] with a 10-fold cross-validation. Based on the resulting model, each pair is then ranked according to the *score* assigned by the classifier, which amounts to a pointwise ranking strategy. Once mentions are locally linked to the top-ranked candidates in all readings, the latter must in turn be ranked in order to finally disambiguate the current text segment. This ranking is done by assigning to each reading the score of its top-ranked candidate or, when the reading contains a sequence of mentions, the minimum of all top-ranked candidates' scores.

On the evaluation data, the system obtains encouraging results of linking accuracy (A) over the set of correctly detected entities (NER columns), compared to the top TAC-KB population evaluation results over English data with only correct entity mentions (table 1). The overall task, i.e. of joint entity recognition and linking, is also measured and is not comparable to equivalent work to the best of our knowledge. Results show that the system's ability to detect false entity mentions should be considerably improved, but filters out some NER noise, which could not be the case in a mere sequential system where all detected mentions would be equally handled by the linking module.

Conducted over a corpus of $\approx$ 400,000 news items of 2009 and 2010 on 16 news topics such as *Politics, Economy* or *Culture*, the EL step results in a set of 46,616 distinct identified entities (35%, 50% and 15% of GPEs, persons and organizations), along with information retrieved from Aleda. Moreover, each entity is associated with new information gathered at each of its occurrence in the extraction corpus. Hence the most frequent association of entities with news features (see *Extracted entities* in fig. 1) augment the extraction list with useful knowledge for further integration.

[6] http://alpage.inria.fr/~sagot/sxpipe.html

[7] Megam: http://www.cs.utah.edu/~hal/megam/

	NER			EL	Joint NER+EL			
	P	R	F	A (all/NILs)	P	R	F	NAE
AFP (French)	0.849	0.768	0.806	0.871/0.597	0.740	0.669	0.702	33%
Top TAC-KBP (English)	-	-	-	0.821/0.891	-	-	-	-

Table 1: NER, linking and joint results

4 Target KB Population and Enrichment

The entities identified during the previous step are made available for the target KB population. Given the specialized use intended for AMO in further content enrichment, a high precision rate is expected w.r.t. its instances, which requires a phase of human intervention. However this process should not come down to a time costly and off-putting task, and should rely on concrete and systematic validation criteria. The extracted entities are presented to domain specialists in order to assess their relevance w.r.t. the KB. This judgement can leverage various type of information, including the usage of entities in the extraction corpus. Each entity is submitted to validation with its Wikipedia or GeoNames URL, which allows for an unambiguous verification of its identity on the Web. Furthermore, candidates are submitted in reverse frequency order: the mention rate of an entity over a given period can by itself indicate its relevance for a KB integration.[8] The validation of less frequent entities can rely in a greater extend on contextual information derived from news features: news topics, keywords, salient words and co-occurring entities can indicate the level of contribution of a given entity to the domain, and therefore its usefulness as a KB instance for further content enrichment. Moreover, the news publication dates collected for each occurrence allow to visualize an entity salience in a given period of time, and can indicate its emergence as an important news actor at certain event peaks.

Following this process, 5,759 relevant entities were selected. Their concrete integration in AMO consists in an automatic instantiation within the underlying ontological model (see AMO in fig. 1). The adequate AMO *Entity* subclass is determined by a straightforward mapping from Aleda entity types to AMO *Entity* subclasses. Finer subclasses, such as *Location/POI* (*point of interest*) can be instantiated based on GeoNames *features* associated with the entity (e.g., locations with the GeoNames *museum* feature are mapped to the *POI* subclass). The instantiation process also considers useful information available on entities: Aleda attributes - the entity normalized name or its geographical coordinates -, are represented in the form of an *owl:dataProperty*. Knowledge elements extracted from the corpus, such as the news topic with which the entity is most mentioned, give rise to *owl:objectProperties* (whose domain and range are the considered entity and an adequate instance of the ontology, e.g. the *Politics* instance of the *NewsTopic* class).

When the process described in section 3 is repeated over new data, i.e. over the daily news flow, and along with regular updates of the Aleda to take into account new entries, the entities presented to validation can either be new w.r.t. AMO or already present as an instance. In the latter case, the additional information linked to the entity should be merged with the adequate existing node. Automatically identifying the adequate existing instance can be achieved by entity resolution techniques, such as systematic comparison of attribute values. In the case of a new entity, a new instance should be created as described above. This illustrates the challenge of dynamic enrichment of resources, comparable to the automatic detection of neologisms in the general language for lexical resources. This aspect of the KB population is intended to be at the center of further developments of our architecture.

In order to integrate AMO in the Linked Data (LD) framework, we applied entity resolution *via* URIs matches with existing LD datasets: all instances defining a Wikipedia URL could be linked to the equivalent DBpedia resource; 20% of AMO instances after the initial population were linked to the NYT data, thus making AMO a suitable resource for content enrichment and Web publishing in the news domain.

[8] Roughly 20% of the distinct entities constitute 80% of the total occurrences in corpus (more than 4 million mentions); hence examining most frequent entities quickly allows for an initial population with the most prominent instances.

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*, 10–18, Suntec, Singapour.

R. Bunescu and M. Pasca. 2006 Using encyclopedic knowledge for named entity disambiguation. In *Proceeding of EACL*, 6:9–16, Trento, Italy.

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 EMNLP-CoNLL*, 708–716, Prague, Czech Republic.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel and R. Weischedel. 2004 The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. In *Proceedings of LREC 2004*, 4:837–840, Lisbon, Portugal

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*, Beijing, China.

H. Ji and R. Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Portland (OR), USA.

H. Ji, R. Grishman and H. Trang Dang. 2011. An Overview of the TAC2011 Knowledge Base Population Track. In *Proceedings of Text Analysis Conference (TAC2011)*,

G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer and R. Lee. 2009. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In *The Semantic Web: Research and Applications*, Springer.

B. Sagot and R. Stern. 2011. Aleda, a free large-scale entity database for French. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Z. Syed, T. Finin and A. Joshi. 2008. Wikipedia as an ontology for describing documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*, 136–144, Seattle (WA), USA.

DBpedia. `http://dbpedia.org/`

GeoNames. Web database:
`http://www.geonames.org/`
Downloadable data:
`http://download.geonames.org/export/dump/`

NYT. The New York Times Linked Data Web site:
`http://data.nytimes.com/`

Wikipedia. French Wikipedia (XML version):
`http://dumps.wikimedia.org/frwiki/`