A Joint Named Entity Recognition and Entity Linking System

Rosa Stern, Benoît Sagot, Frédéric Béchet

A Joint Named Entity Recognition and Entity Linking System

Rosa Stern,[1,2] **Benoît Sagot**[1] **and Frédéric Béchet**[3]
[1]Alpage, INRIA & Univ. Paris Diderot, Sorbonne Paris Cité / F-75013 Paris, France
[2]AFP-Medialab / F-75002 Paris, France
[3]Univ. Aix Marseille, LIF-CNRS / Marseille, France

Abstract

We present a joint system for named entity recognition (NER) and entity linking (EL), allowing for named entities mentions extracted from textual data to be matched to uniquely identifiable entities. Our approach relies on combined NER modules which transfer the disambiguation step to the EL component, where referential knowledge about entities can be used to select a correct entity reading. Hybridation is a main feature of our system, as we have performed experiments combining two types of NER, based respectively on symbolic and statistical techniques. Furthermore, the statistical EL module relies on entity knowledge acquired over a large news corpus using a simple rule-base disambiguation tool. An implementation of our system is described, along with experiments and evaluation results on French news wires. Linking accuracy reaches up to 87%, and the NER f-measure up to 83%.

1 Introduction

1.1 Textual and Referential Aspects of Entities

In this work we present a system designed for the extraction of entities from textual data. Named entities (NEs), which include person, location, company or organization names[1] must therefore be detected using named entity recognition (NER) techniques. In addition to this detection based on their surface forms, NEs can be identified by mapping them to the actual entity they denote, in order for these extractions to constitute useful and complete information. However, because of name *variation*, which can be surfacic or encyclopedic, an entity can be denoted by several *mentions* (e.g., *Bruce Springsteen*, *Springsteen*, *the Boss*); conversely, due to name *ambiguity*, a single mention can denote several distinct entities (*Orange* is the name of 22 locations in the world; in French, *M. Obama* can denote both the US president *Barack Obama* (*M.* is an abbreviation of *Monsieur* 'Mr') or his spouse *Michelle Obama*; in this case ambiguity is caused by variation). Even in the case of unambiguous mentions, a clear link should be established between the surface mention and a uniquely identifiable entity, which is achieved by entity linking (EL) techniques.

1.2 Entity Approach and Related Work

In order to obtain referenced entities from raw textual input, we introduce a system based on the joint application of named entity recognition (NER) and entity linking (EL), where the NER output is given to the linking component as a set of possible mentions, preserving a number of ambiguous readings. The linking process must thereafter evaluate which readings are the most probable, based on the most likely entity matches inferred from a similarity measure with the context.

NER has been widely addressed by symbolic, statistical as well as hybrid approaches. Its major part in information extraction (IE) and other NLP applications has been stated and encouraged by several editions of evaluation campaigns such as MUC (Marsh and Perzanowski, 1998), the CoNLL-2003 NER shared task (Tjong Kim Sang and De Meulder, 2003) or ACE (Doddington et al., 2004), where NER systems show near-human performances for the English language. Our system aims at benefitting from both symbolic and statistical NER techniques, which have proven efficient but not necessarily over the same type

[1]The set of possible named entities varies from restrictive, as in our case, to wide definitions; it can also include dates, event names, historical periods, etc.

of data and with different precision/recall tradeoff. NER considers the surface form of entities; some type disambiguation and name normalization can follow the detection to improve the result precision but do not provide referential information, which can be useful in IE applications. EL achieves the association of NER results with uniquely identified entities, by relying on an entity repository, available to the extraction system and defined beforehand in order to serve as a target for mention linking. Knowledge about entities is gathered in a dedicated knowledge base (KB) to evaluate each entity's similarity to a given context. After the task of EL was initiated with Wikipedia-based works on entity disambiguation, in particular by Cucerzan (2007) and Bunescu and Pasca (2006), numerous systems have been developed, encouraged by the TAC 2009 KB population task (McNamee and Dang, 2009). Most often in EL, Wikipedia serves both as an entity repository (the set of articles referring to entities) and as a KB about entities (derived from Wikipedia infoboxes and articles which contain text, metadata such as categories and hyperlinks). Zhang et al. (2010) show how Wikipedia, by providing a large annotated corpus of linked ambiguous entity mentions, pertains efficiently to the EL task. Evaluated EL systems at TAC report a top accuracy rate of 0.80 on English data (McNamee et al., 2010).

Entities that are unknown to the reference database, called *out-of-base* entities, are also considered by EL, when a given mention refers to an entity absent from the available Wikipedia articles. This is addressed by various methods, such as setting a threshold of minimal similarity for an entity selection (Bunescu and Pasca, 2006), or training a separate binary classifier to judge whether the returned top candidate is the actual denotation (Zheng et al., 2010). Our approach of this issue is closely related to the method of Dredze et al. in (2010), where the *out-of-base* entity is considered as another entry to rank.

Our task differs from EL configurations outlined previously, in that its target is entity extraction from raw news wires from the news agency Agence France Presse (AFP), and not only linking relying on gold NER annotations: the input of the linking system is the result of an automatic NER step, which will produce errors of various kinds. In particular, spans erroneously detected as NEs will have to be discarded by our EL system. This case, which we call *not-an-entity*, contitute an additional type of special situations, together with *out-of-base* entities but specific to our setting. This issue, as well as others of our task specificities, will be discussed in this paper. In particular, we use resources partially based on Wikipedia but not limited to it, and we experiment on the building of a domain specific entity KB instead of Wikipedia.

Section 2 presents the resources used throughout our system, namely an entity repository and an entity KB acquired over a large corpus of news wires, used in the final linking step. Section 3 states the principles on which the NER components of our system relies, and introduces the two existing NER modules used in our joint architecture. The EL component and the methodology applied are presented in section 4. Section 5 illustrates this methodology with a number of experiments and evaluation results.

2 Entity Resources

Our system relies on two large-scale resources which are very different in nature:

- the entity database Aleda, automatically extracted from the French Wikipedia and `Geonames`;
- a knowledge base extracted from a large corpus of AFP news wires, with distributional and contextual information about automatically detected entites.

2.1 Aleda

The Aleda entity repository[2] is the result of an extraction process from freely available resources (Sagot and Stern, 2012). We used the French Aleda databased, extracted the French Wikipedia [3] and `Geonames` [4]. In its current development, it provides a generic and wide coverage entity resource accessible *via* a database. Each entity in Aleda is associated with a range of attributes, either referential (e.g., the type of the entity among *Person*, *Location*, *Organization* and *Company*, the population for a location or the gender of a person, etc.) or formal, like the entity's URI from

[2] Aleda is part of the Alexina project and freely available at `https://gforge.inria.fr/projects/alexina/`.

[3] `www.fr.wikipedia.org`

[4] `www.geonames.org`

Wikipedia or Geonames; this enables to uniquely identify each entry as a Web resource.

Moreover, a range of possible *variants* (*mentions* when used in textual content) are associated to entities entries. Aleda's variants include each entity's canonical name, Geonames location labels, Wikipedia redirection and disambiguation pages aliases, as well as dynamically computed variants for person names, based in particular on their first/middle/last name structure. The French Aleda used in this work comprises 870,000 entity references, associated with 1,885,000 variants.

The main informative attributes assigned to each entity in Aleda are listed and illustrated by examples of entries in Tab. 1. The popularity attribute is given by an approximation based on the length of the entity's article or the entity's population, from Wikipedia and Geonames entries respectively. Table 1 also details the structure of Aleda's variants entries, each of them associated with one or several entities in the base.

Unlike most EL systems, Wikipedia is not the entity base we use in the present work; rather, we rely on the autonomous Aleda database. The collect of knowledge about entities and their usage in context will also differ in that our target data are news wires, for which the adaptability of Wikipedia can be questioned.

2.2 Knowledge Acquisition over AFP news

The linking process relies on knowledge about entities, which can be acquired from their usage in context and stored in a dedicated KB. AFP news wires, like Wikipedia articles, have their own structure and formal metadata: while Wikipedia articles each have a title referring to an entity, object or notion, a set of *categories*, hyperlinks, etc., AFP news wires have a headline and are tagged with a *subject* (such as *Politics* or *Culture*) and several *keywords* (such as *cinema*, *inflation* or *G8*), as well as information about the date, time and location of production. Moreover, the distribution of entities over news wires can be expected to be significantly different from Wikipedia, in particular w.r.t. uniformity, since a small set of entities forms the majority of occurrences. Our particular context can thus justify the need for a domain specific KB.

As opposed to Wikipedia where entities are identifiable by hyperlinks, AFP corpora provide no such indications. Wikipedia is in fact a corpus where entity mentions are clearly and uniquely linked, whereas this is what we aim at achieving over AFP's raw textual data. The acquisition of domain specific knowledge about entities from AFP corpora must circumvent this lack of indications. In this perspective we use an implementation of a *naive linker* described in (Stern and Sagot, 2010). For the main part, this system is based on heuristics favoring popular entities in cases of ambiguities. An evaluation of this system showed good accuracy of entity linking (0.90) over the subset of correctly detected entity mentions:[5] on the evaluation data, the resulting NER reached a precision of 0.86 and a recall of 0.80. Therefore we rely on the good accuracy of this system to identify entities in our corpus, bearing in mind that it will however include cases of false detections, while knowledge will not be available on missed entities. It can be observed that by doing so, we aim at performing a form of co-training of a new system, based on supervised machine learning. In particular, we aim at providing a more portable and systematic method for EL than the heuristics-based naive linker which is highly dependent on a particular NER system, SXPipe/NP, described later on in section 3.2.

The knowledge acquisition was conducted over a large corpus of news wires (200,000 news items of the years 2009, 2010 and part of 2011). For each occurrence of an entity identified as such by the naive linker, the following features are collected, updated and stored in the KB at the entity level: (i) entity total occurrences and occurrences with a particular mention; (ii) entity occurrence with a news item topics and keywords, most salient words, date and location; (iii) entity co-occurrence with other entity mentions in the news item. These features are collected for both entities identified by the naive linker as Aleda's entities and mentions recognized by NER pattern based rules; the latter account for out-of-base entities, approximated by a cluster of all mentions whose normalization returns the same string. For instance, if the mentions *John Smith* and *J. Smith* were detected in a document but not linked to an entity in Aleda, it would be assumed that they co-refer to an entity whose normalized

[5] This subset is defined by a strict span and type correct detection, and among the sole entities for which a match in Aleda or outside of it was identified; the evaluation data is presented in section 5.1.

Entities

ID	Type	CanonicalName	Popularity	URI
20013	Loc	Kingdom of Spain	46M	geon:2510769
10063	Per	Michael Jordan	245	wp:Michael_Jordan
20056	Loc	Orange (California)	136K	geon:5379513
10039	Comp	Orange	90	wp:Orange_(entreprise)

Variants

ID	Variant	FirstName	MidName	LastName
20013	Espagne	–	–	–
10063	Jordan	–	–	Jordan
10029	George Walker Bush	George	Walker	Bush
10039	Orange	–	–	–
20056	Orange	–	–	–

Table 1: Structure of Entities Entries and Variants in Aleda

name would be *John Smith*; this *anonymous entity* would therefore be stored and identified *via* this normalized name in the KB, along with its occurrence information.

3 NER Component

3.1 Principles

One challenging subtask of NER is the correct detection of entity mentions *spans* among several ambiguous readings of a segment. The other usual subtask of NER consists in the labeling or classification of each identified mention with a *type*; in our system, this functionality is used as an indication rather than a final attribute of the denoted entity. The type assigned to each mention will in the end be the one associated with the matching entity. The segment *Paris Hilton* can for instance be split in two consecutive entity mentions, *Paris* and *Hilton*, or be read as a single one. Whether one reading or the other is more likely can be inferred from knowledge about entities possibly denoted by each of these three mentions: depending on the considered document's topic, it can be more probable for this segment to be read as the mention *Paris Hilton*, denoting the celebrity, rather than the sequence of two mentions denoting the capital of France and the hotel company. Based on this consideration, our system relies on the ability of the NER module to preserve multiple readings in its output, in order to postpone to the linker the appropriate decisions for ambiguous cases. Two NER systems fitted with this ability are used in our architecture.

Figure 1: Ambiguous NER output for the segment *Paris Hilton* in SXPipe/NP

3.2 Symbolic NER: SXPipe/NP

NP is part of the SXPipe surface processing chain (Sagot and Boullier, 2008). It is based on a series of recognition rules and on a large coverage lexicon of possible entity variants, derived from the Aleda entity repository presented in section 2.1. As an SXPipe component, NP formalizes the text input in the form of directed acyclic graphs (DAGs), in which each possible entity mention is represented as a distinct transition, as illustrated in Figure 1. Possible mentions are labeled with *types* among *Person, Location, Organization and Company*, based on the information available about the entity variant in Aleda and on the type of the rule applied for the recognition.

Figure 1 also shows how an alternative transition is added to each mention reading of a segment, in order to account for a possible non-entity reading (i.e., for a *false match* returned by the NER module). When evaluating the adequacy of each reading, the following EL module will in fact consider a special *not-an-entity* candidate as a possible match for each mention, and select it as the most probable if competing entity readings prove insufficiently adequate w.r.t. the considered context.

3.3 Statistical NER: LIANE

The statistical NER system LIANE (Bechet and Charton, 2010) is based on (i) a generative HMM-based process used to predict part-of-speech and semantic labels among *Person, Location, Organi-*

zation and Product for each input word[6], and (ii) a discriminative CRF-based process to determine the entity mentions' spans and overall type. The HMM and CRF models are learnt over the ESTER corpus, consisting in several hundreds of hours of transcribed radio broadcast (Galliano et al., 2009), annotated with the BIO format (table 2). The out-

investiture	NFS	O
aujourd'hui	ADV	B-TIME
à	PREPADE	O
Bamako	LOC	B-LOC
Mali	LOC	B-LOC

Table 2: BIO annotation for LIANE training

put of LIANE consists in an n-best lists of possible entity mentions, along with a confidence score assigned to each result. Therefore it also provides several readings of some text segments, with alternatives of entity mention readings.

As shown in (Bechet and Charton, 2010), the learning model of LIANE makes it particularly robust to difficult conditions such as non capitalization and allows for a good recall rate on various types of data. This is in opposition with manually handcrafted systems such as SXPipe/NP, which can reach high precision rates over the development data but prove less robust otherwise. These considerations, as well as the benefits of a cooperations between these two types of systems are explored in (Béchet et al., 2011).

By coupling LIANE and SXPipe/NP to perform the NER step of our architecture, we expect to benefit from each system's best predictions and improving the precision and recall rates. This is achieved by not enforcing disambiguation of spans and types at the NER level but by transferring this possible source of errors to the linking step, which will rely on entity knowledge rather than mere surface forms to determine the best readings, along with the association of mentions with entity references.

4 Linking Component

4.1 Methodology for Best Reading Selection

As previously outlined, the purpose of our joint architecture is to infer best entity readings from

[6]For the purpose of type consistency across both NER modules, the NP type *Company* is merged with *Organization*, and the LIANE mentions typed as *Product* are ignored since they are not yet supported by the overall architecture.

Figure 2: Possible readings of the segment *Paris Hilton* and ordered candidates

contextual similarity between entities and documents rather than at the surface level during NER. The linking component will therefore process ambiguous NER outputs in the following way, illustrated by Fig. 2.

1. For each mention returned by the NER module, we aim at finding the best fitting entity w.r.t. the context of the mention occurrence, i.e., at the document level. This results in a list of candidate entities associated with each mention. This candidates set always includes the *not-an-entity* candidate in order to account for possible false matches returned by the NER modules.

2. The list of candidates is ordered using a pointwise ranking model, based on the maximum entropy classifier `megam`.[7] The best scored candidate is returned as a match for the mention; it can be either an entity present in Aleda, i.e., a *known* entity, or an *anonymous* entity, seen during the KB acquisition but not resolved to a known reference and identified by a normalized name, or the special *not-an-entity* candidate, which discards the given mention as an entity denotation.

3. Each reading is assigned a score depending on the best candidates' scores in the reading.

The key steps of this process are the selection of candidates for each mention, which must reach a sufficient recall in order to ensure the reference resolution, and the building of the feature vector for each mention/entity pair, which will be evaluated by the candidate ranker to return the most adequate entity as a match for the mention. Throughout this process, the issues usually raised by EL must be considered, in particular the ability for the model to learn cases of *out-of-base* entities, which our system addresses by forming a set of candidates not only from the entity reference base (i.e., Aleda), but also from the dedicated KB where anonymous entities are also collected. Furthermore, unlike the general configuration of EL tasks, such as the TAC KB population task (sec-

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

tion 1.2), our input data does not consist in mentions to be linked but in multiple possibilities of mention readings, which adds to our particular case the need to identify false matches among the queries made to the linker module.

4.2 Candidates Selection

For each mention detected in the NER output, the mention string or *variant* is sent as a query to the Aleda database. Entity entries associated with the given variant are returned as candidates. The set of retrieved entities, possibly empty, constitutes the candidate set for the mention. Because the knowledge acquisition included the extraction of unreferenced entities identified by normalized names (section 2.2), we can send the normalization of the mention as an additional query to our KB. If a corresponding anonymous entity is returned, we can create an *anonymous* candidate and add it to the candidate set. *Anonymous* candidates account for the possibility of an *out-of-base* entity denoted by the given mention, with respectively some and no information about the potential entity they might stand for. Finally, the set is augmented with the special *not-an-entity* candidate.

4.3 Features for Candidates Ranking

For each pair formed by the considered mention and each entity from the candidate set, we compute a feature vector which will be used by our model for assessing the probability that it represents a correct mention/entity linking. The vector contains attributes pertaining to the mention, the candidate and the document themselves, and to the relations existing between them.

Entity attributes Entity attributes present in Aleda and the KB are used as features: Aleda provides the entity type, a popularity indication and the number of variants associated with the entity. We retrieve from the KB the entity frequency over the corpus used for knowledge acquisition.

Mention attributes At the mention level, the feature set considers the absence or presence of the mention as a variant in Aleda (for any entity), its occurrence frequency in the document, and whether similar variants, possibly indicating name variation of the same entity, are present in the document (similar variants can have a string equal to the mention's string, longer or shorter than the mention's string, included in the mention's string or including it). In the case of a mention returned by LIANE, the associated confidence score is also included in the feature set.

Entity/mention relation The comparison between the surface form of the entity's canonical name and the mention gives a similarity rate feature. Also considered as features are the relative occurrence frequency of the entity w.r.t. the whole candidate set, the existence of the mention as a variant for the entity in Aleda, the presence of the candidate's type (retrieved from Aleda) in the possible mention types provided by the NER. The KB indicates frequency of its occurrences with the considered mention, which adds another feature.

Document/entity similarity Document metadata (in particular topics and keywords) are inherited by the mention and can thus characterize the entity/mention pair. Equivalent information was collected for entities and stored in the KB, which allows to compute a cosine similarity between the document and the candidate. Moreover, the most salient words of the document are compared to the ones most frequently associated with the entity in the KB. Several atomic and combined features are derived from these similarity measures.

Other features pertain to the NER output configuration, as well as possible false matches:

NER combined information One of the two available NER modules is selected as the base provider for entity mentions. For each mention which is also returned by the second NER module, a feature is instanciated accordingly.

Non-entity features In order to predict cases of *not-an-entity* readings of a mention, we use a generic lexicon of French forms (Sagot, 2010) where we check for the existence of the mention's variant, both with and without capitalization. If the mention's variant is the first word of the sentence, this information is added as a feature.

These features represent attributes of the entity/mention pair which can either have a boolean value (such as variant presence or absence in Aleda) or range throughout numerical values (e.g., entity frequencies vary from 0 to 201,599). In the latter case, values are discretized. All features in our model are therefore boolean.

4.4 Best Candidate Selection

Given the feature vector instanciated for an (candidate entity, mention) pair, our model assigns it a score. All candidates in the subset are then ranked accordingly and the first candidate is returned as

the match for the current mention/entity linking. *Anonymous* and *not-an-entity* candidates, as defined earlier and accounting respectively for potential *out-of-base* entity linking and NER false matches, are included in this ranking process.

4.5 Ranking of Readings

The last step of our task consists in the ranking of multiple readings and has yet to be achieved in order to obtain an output where entity mentions are linked to adequate entities. In the case of a reading consisting in a single transition, i.e., a single mention, the score is equal to the best candidate's score. In case of multiple transitions and mentions, the score is the minimum among the best candidates' scores, which makes a low entity match probability in a mention sequence penalizing for the whole reading. Cases of false matches returned by the NER module can therefore be discarded as such in this step, if an overall non-entity reading of the whole path receives a higher score than the other entity predictions.

5 Experiments and Evaluation

5.1 Training and Evaluation Data

We use a gold corpus of 96 AFP news items intended for both NER and EL purposes: the manual annotation includes mention boundaries as well as an entity identifier for each mention, corresponding to an Aleda entry when present or the normalized name of the entity otherwise. This allows for the model learning to take into account cases of *out-of-base* entities. This corpus contains 1,476 mentions, 437 distinct Aleda's entries and 173 entities absent from Aleda. All news items in this corpus are dated May and June 2009.

In order for the model to learn from cases of *not-an-entity*, the training examples were augmented with false matches from the NER step, associated with this special candidate and the positive class prediction, while other possible candidates were associated with the negative class. Using a 10-fold cross-validation, we used this corpus for both training and evaluation of our joint NER and EL system.

It should be observed that the learning step concerns the ranking of candidates for a given mention and context, while the final purpose of our system is the ranking of multiple readings of sentences, which takes place after the application of our ranking model for mention candidates. Thus our system is evaluated according to its ability to choose the right reading, considering both NER recall and precision and EL accuracy, and not only the latter.

5.2 Task Specificities

As outlined in section 1.2, the input for the standard EL task consists in sets of entity mentions from a number of documents, sent as queries to a linking system. Our current task differs in that we aim at both the extraction and the linking of entities in our target corpus, which consists in unannotated news wires. Therefore, the results of our system are comparable to previous work when considering a setting where the NER output is in fact the gold annotation of our evaluation data, i.e., when all mention queries should be linked to an entity. Without modifying the parameters of our system (i.e., no deactivation of false matches predictions), we obtain an accuracy of 0.76, in comparison with a TAC top accuracy of 0.80 and a median accuracy of 0.70 on English data.[8]

It is important to observe that our data consists only in journalistic content, as opposed to the TAC dataset which included various types of corpora. This difference can lead to unequally difficulty levels w.r.t. the EL task, since NER and EL in journalistic texts, and in particular news wires, tend to be easier than on other types of corpora. This comes among other things from the fact that a small number of popular entities constitute the majority of NE mention occurrences.

In most systems, EL is performed over noisy NER output and participates to the final decisions about NEs extractions. Therefore the ability of our system to correctly detect entity mentions in news content is estimated by computing its precision, recall and f-measure.[9] The EL accuracy, i.e., the rate of correctly linked mentions, is mea-

[8] As explained previously, these figures, as well as the ones presented later on, cannot be compared with the 0.90 score obtained by the naive linker which we used for the entity KB acquisition. This score is obtained only on mentions identified by the SXPipe/NP system with the correct span and type, whereas our system does not consider the mention type as a contraint for the linking process, and on correct identification of a match in or outside of Aleda.

[9] Only mention boundaries are considered for NER evaluation, while other settings require correct type identification for validating a fully correct detection. In our case, NER is not a final step, and entity typing is derived from the entity linking result.

Setting	NER			EL	Joint NER+EL		
	Precision	Recall	f-measure	Accuracy	Recall	Precision	f-measure
SXPipe/NP	0.849	0.768	0.806	*0.871*	0.669	0.740	0.702
LIANE	0.786	0.891	0.835	*0.820*	0.730	0.645	0.685
SXPipe/NP- NL	0.775	0.726	0.750	*0.875*	0.635	0.678	0.656
LIANE- NL	0.782	0.886	0.831	*0.818*	0.725	0.640	0.680
SXPipe/NP & 2	0.812	0.747	0.778	*0.869*	0.649	0.705	0.676
LIANE & SXPipe/NP	0.803	0.776	0.789	*0.859*	0.667	0.689	0.678

Table 3: Joint NER and EL results. *Each* EL *accuracy covers a different set of correctly detected mentions*

sured over the subset of mentions whose reading was adequately selected by the final ranking. The evaluation of our system has been conducted over the corpus described previously with settings presented in the next section.

5.3 Settings and results

We used each of the two available NER modules as a provider for entity mentions, either on its own or together with the second system, used as an indicator. For each of these settings, we tried a modified setting in which the prediction of the naive linker (NL) used to build the entity KB (section 2.2) was added as a feature to each mention/candidate pair (settings SXPipe/NP-NL and LIANE-NL). These experiments' results are reported in Table 3 and are given in terms of:

- NER precision, recall and f-measure;
- EL accuracy over correctly recognized entities; therefore, the different figures in column EL Accuracy are not directly comparable to one another, as they are not obtained over the same set of mentions;
- joint NER+EL precision, recall and f-measure; the precision/recall is computed as the product of the NER precision/recall by the EL accuracy.

As expected, SXPipe/NP performs better as far as NER precision is concerned, and LIANE performs better as far as NER recall is concerned. However, the way we implemented hybridation at the NER level does not seem to bring improvements. Using the output of the naive linker as a feature leads to similar or slightly lower NER precision and recall. Finally, it is difficult to draw clear-cut comparative conclusions at this stage concerning the joint NER +EL task.

6 Conclusion and Future Work

We have described and evaluated various settings for a joint NER and EL system which relies on the NER systems SXPipe/NP and LIANE for the NER step. The EL step relies on a hybrid model, i.e., a statistical model trained on a manually annotated corpus. It uses features extracted from a large corpus automatically annotated and where entity disambiguations and matches were computed using a basic heuristic tool. The results given in the previous section show that the joint model allows for good NER results over French data. The impact of the hybridation of the two NER modules over the EL task should be further evaluated. In particular, we should investigate the situations where an mention was incorrectly detected (e.g., the span is not fully correct) although the EL module linked it with the correct entity. Moreover, a detailed evaluation of out-of-base linkings vs. linking in Aleda remains to be performed.

In the future, we aim at exploring various additional features in the EL system, in particular more combinations of the current features. The adaptation of our learning model to NER combinations should also be improved. Finally, a larger set of training data should be considered. This shall become possible with the recent manual annotation of a half-million word French journalistic corpus.

References

F. Bechet and E Charton. 2010. Unsupervised knowledge acquisition for extracting named entities from speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

F. Béchet, B. Sagot, and R. Stern. 2011. Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un

système d'étiquetage en entités nommées. In *Actes de la Conférence TALN 2011*, Montpellier, France.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL*, volume 2007, pages 708–716.

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 - Volume 4*, pages 837–840.

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*, pages 277–285.

S. Galliano, G. Gravier, and L. Chaubard. 2009. The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *Interspeech 2009*.

E. Marsh and D. Perzanowski. 1998. Muc-7 evaluation of ie technology: Overview of results. In *Proceedings of the Seventh Message Understanding Conference (MUC-7) - Volume 20*.

P. McNamee and H.T. Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*.

P. McNamee, H.T. Dang, H. Simpson, P. Schone, and S.M. Strassel. 2010. An evaluation of technologies for knowledge base population. *Proc. LREC2010*.

B. Sagot and P. Boullier. 2008. SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, 49(2):155–188.

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

B. Sagot. 2010. The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, Vallette, Malta.

R. Stern and B. Sagot. 2010. Détection et résolution d'entités nommées dans des dépêches d'agence. In *Actes de la Conférence TALN 2010*, Montréal, Canada.

E. F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*, pages 142–147, Edmonton, Canada.

W. Zhang, J. Su, C.L. Tan, and W.T. Wang. 2010. Entity linking leveraging: automatically generated annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1290–1298.

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*, pages 483–491.