Resources for Named Entity Recognition and Resolution in News Wires

Rosa Stern[1,2] **and Benoît Sagot**[1]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 30 rue du Château des Rentiers, 75013 Paris, France
2. Agence France-Presse – Medialab, 2 place de la Bourse, 75002 Paris, France
rosa.stern@afp.com, benoit.sagot@inria.fr

Abstract

In the applicative context of news wire enrichment with metadata, named entity recognition plays an important role, but requires to be followed by a resolution module that maps named entity mentions to entries in a reference database. In this paper, we describe NP, the named entity module embedded in the SXPipe shallow processing chain, that we used for extracting information from French news wires from the Agence France-Presse. We describe the construction of our reference database from freely available external resources, as well as our named entity detection, disambiguation and resolution modules. We also introduce a freely available and manually developped annotated corpus designed for the evaluation of named entity recognition and resolution tools, and provide evaluation figures for NP.

1. Overview

1.1. Introduction

Reference resolution is a challenge any named entity recognition system (henceforth NER) is confronted with as soon as its integration within the frame of a real application is in question. As studies like [Blume, 2005] have stated, reference resolution is obviously beneficial, necessary indeed, for an application involving NER in the prospect of exploiting information provided by named entities (henceforth NE) detected throughout data. Identification of certain text segments as NE phrases doesn't bring the whole information conveyed by NE usage without this segment — or mention — being linked to an extra-textual reference. Any particular application needing this association assumes the existence of such references, or more precisely of a reference resource to be the target of this linking.

After reviewing some considerations which are of importance in the development of an NER and resolution system, we present the application context of our research, which involves a dedicated reference base (section 2.). The creation and maintenance of the latter depends on the performance and functionalities of the system described in section 3., which addresses both entity recognition and reference resolution in a modular yet integrated fashion, applied to data in French. While the NER module (section 3.2.) involves some features which are peculiar to French and must be adapted in order to handle other languages, the resolution module presented in 3.3. is not language specific and is portable in the different steps of the system's evolution. Our system's development has lead to the creation of an evaluation resource, presented in section 5. along with results and future work outline.

1.2. Reference ambiguity

As done in the frame of ACE shared task [ACE08, 2008], NE mentions can be resolved to a unique reference through an identifier which disambiguate any linguistic metonymy or variation phenomena.

Entity resolution indeed consists of assigning a reference to phrases detected in texts as named entities mentions. In order to address this task within an NER system, the limits of classical entity classifications through static type categories have to be stressed (as in [Poibeau, 2005] or [Murgatroyd, 2008]). Such limits concern cases like metonymy ("I've read Proust and didn't like it"), precise oganization typing ("The UN will act on this violation"), entity inclusion in complex entities ("Kennedy Airport"). These cases illustrate the similarity between the polysemic behaviours of NEs and terms [Brun et al., 2009b]. The handling of entity reference in any NER system therefore implies the integration of extra-linguistic values of NEs, which is not necessarily obvious in their linguistic usage. Those characteristics of NE mentions make entity reference more subtle than a one-to-one matching between a detected phrase in texts and an identifier. In particular, the typing of NE mentions at the textual level must both guide the resolution process and avoid confusion. In the case of "Proust" in the above example, the fact that this particular mention doesn't refer to the person identified as the French writer *Marcel Proust* but, by metonymy, to his work raises the issue of the actual reference of this mention. Whether it should be resolved to the person *Marcel Proust* or not is a decision to be made in the context of the particular final application, but the linking of this mention to this reference must at least be detected and stated by the system. In the case of "Kennedy Airport", it is more obvious that the correct typing of this entity mention (type *facility*) should prevent any reference resolution to the person *J.F. Kennedy*. NE mentions found in texts can thus be treated as ambiguous occurences which an appropriate tool, such as described in section 3.3., can resolve to an identifier after a more general detection phase by taking into account several relevant elements given by the text about the extra-linguistic context of those occurences.

The other challenging aspect of entity reference, not without many connexions with the problem of polysemy,

is the multiplicity of ways available to refer to a particular entity on a linguistic and textual level (i.e. a form of synonymy). Whether graphic or grammatical variants are in question (as "Mr Obama", "B. Obama" or "He", "The US president" for Barack Obama), such configurations raise the issue of an obvious matching between a mention of a NE and an identifier. Whereas the case of graphic variants are relatively easy to predict and handle within a surface analysis based tool, the grammatical anaphoras and coreferences demand to be resolved at a a deeper and more complex analysis level.

2. Application Context

2.1. Information Enrichment with Metadata

Alpage and the Agence France Presse (AFP) Medialab department have been involved in several projects dealing with information extraction from news wires and enrichment of AFP's production. One of the main prospect of this research is to build a general and standardized reference base of metadata used in the production of news. By indexing news wires according to this resource, a structured and coherent production enrichment would be possible. This would then help for the improvment of various applications specific to a press agency, such as the filtering of the news production with customer-specific parameters or documentation tasks through information research in archived news.

The NER system which we present here has already been integrated to the SAPIENS platform, which is a prototype of news wires exploitation for information retrieval and knowledge extraction, on the particular topic of quotations. SAPIENS allows for a full quotations detection by matching them with their author, i.e. NEs mentioned in news wires. The user can therefore select some name among the list of detected entities and consult the quotations made by the corresponding person, all grouped together or in the context of the original news wire and of other entities related to them.

2.2. Workflow

This information enrichment with metadata has two aspects: it involves on the one hand the detection of NE as they are considered a decisive kind of metadata for the description of news content; on the other hand, not the entire set of metadata detected is considered relevant for a reference base whose main purpose is to reflect a state of knowledge inferable from the news production. As such, entities detected in news texts are always relevant to the description of the particular item they were found in, whereas they are not always relevant to the general and high-level reference base. NER in this context must therefore handle reference resolution at two levels. The detection tool itself has access to various databases, designed for NER and containing references; the matching of detected entities to references in those databases is then followed by a lookup in the reference base. At this point it must check the obtained reference against the reference base records and do one of these two actions: link the entity to its matching record if it exists, or propose the entity as a candidate to the reference base. The NER system thus updates its resources every time it operates this confrontation: an entity which is actually a record of the reference base should never be skipped, whereas a candidate entity rejected as a reference base record must not be reevaluated each time it occurs in texts . This information about entities reference must be taken into account by the NER system and passed on to its resources, which thus evolve along with their confrontation to new data. Reference resolution in our project thus happens at two levels: the level of the NER system resources, and the level of the reference base designed for the knowledge description of the AFP production. This modularity widely influences the type of resources used by the NER system, i.e. they have to include a set of entities references as large, relevant and exhaustive as possible in order to propose adequate candidates to the matching with the reference base. The choice and usage of those resources is described in section 3.1. It can be noted that among the different types of entities detected, some give more rise to reference resolution than others, i.e. *Persons* and *Organizations* show more peculiarities regarding resolution tasks as outlined in introduction than *Locations*.

In the particular case of a press agency, reference resolution takes on further complexities which have to be addressed, the first of which being the dynamic nature of the data. As news wires are produced every day (800 in French and a total of 5000 in the six languages in use at the AFP), the information they report and convey by definition introduces new references as new events happen, involving existant and new entities. The latter are either entities which have no previous record in references resources, or which do have one but were not considered as relevant entries for the reference base. Both configurations must be addressed, in the first case by proposing new references to handle unknown entities and in the second by promoting the entity status to the one of candidate for the reference base.

2.3. Reference Base

The building of the reference base is driven by a description of the data it should represent, i.e. a conceptual description of the knowledge reported and conveyed by the news production. This conceptual description takes the form of an ontology whose core is the relation network organizing entities. The entities classes outlined by this ontology correspond to a certain extent to a usual entities typology, mainly to a classification among *Person*, *Organization* (including companies, institutional organizations, sport teams...) and *Location* (including geopolitical entities as well as points of interests or facilities). This conceptual model for the reference base is also set to reflect the news themes already in use as metadata at the AFP and directly ensued by the IPTC taxonomy [1]. Politics, Economy, Culture or Sport are such themes and come along with a series

[1] http://www.iptc.org/

of subcategories used to describe news content. Those will have to be integrated in the network formed by the entities found in texts and populating the reference base.

The reference base in itself is currently in development. The phase consisting in matching detected entities against this base and in updating NER resources accordingly is therefore not fully realized yet in the running of our system. However, this preliminary usage of the system is not isolated from the overall application. It will indeed be used to build a first set of records to be integrated to the base, before next iterations where the lookup and candidacy phases take place.

The integration of NE recognition and resolution in this application therefore allows for the reference base population, as well as for its maintenance: the news production will be processed by the NER system combined with other specialized modules; each news item will then be indexed according to the data extracted by the system, depending on their mapping with the reference base.

3. NP: a system for NER and Entity Reference Disambiguation

Our NER and Entity Reference Disambiguation system is a part of SXPipe, a robust and modular surface processing chain for various languages and unrestricted text, used for shallow analysis or pre-preprocessing before parsing [Sagot and Boullier, 2005; Sagot and Boullier, 2008]. SXPipe is a freely-available[2] set of tools which performs (1) "named entities" recognition: pre-tokenization named entities (URLs, emails, dates, addresses, numbers...), (2) tokenization and segmentation in sentences, (3) token-based named entities (phrases in foreign languages...), (4) non-deterministic multi-word units detection and spelling error correction, and (5) lexicon-based patterns detection. The NE module within SXPipe, called NP (from the French *Noms Propres*, "Proper Nouns"), belongs to the 5th step.

NP is divided in two different steps described below. The first step is a non-deterministic detection module, developed in SXPipe's `dag2dag` framework [Sagot and Boullier, 2008]. This framework allows for defining context-free patterns and for using dedicated gazetteers, while remaining very efficient in terms of processing time. The second step disambiguates the ambiguous output of the first step and resolves the NEs that are retained, w.r.t. a NE database described below. These two steps need not be consecutive. The rationale behind this is that other modules that are applied between NE detection and NE disambiguation/normalization could achieve some disambiguation. For example, in the above-described SAPIENS system for quotation detection, the verbatim quotation detection module chooses the type *Person* when a NE is interpreted as the author of a verbatim quotation, even when it is competing with other types (e.g., *Marseille*, which is both a city and the last name of a mediatic historian and economist).[3]

Both modules rely on a large NE database extracted from several large-scale information repositories, mostly Wikipedia and `GeoNames`, as we shall now describe.

Type	n# of entries	n# of variants
Person	263,035	883,242
Location	551,985	624,175
Organization	17,807	44,983
Work	27,222	59,851
Company	9,000	17,252
Product	3,648	6,350

Table 1: Quantitative data about the NE database underlying NP. Note that we have focused mostly on person, organization and location names.

3.1. Resources and NE database building

Our NE database contains almost 900 000 entries and approx. 1,6 million NE denotation variants. It contains location, organization, person, company, product and work (booktitle, movie title...) names. More detailed quantitative information is shown in Table 1.

We extracted this database from two different sources: `GeoNames` for location names,[4] and the French Wikipedia for the other types of NE.[5] Each entry has a unique id, either a `GeoNames` id or an URL pointing to a Wikipedia article.

For location names, we filtered the `GeoNames` database using criteria defined according to the nature of the corpus. Because of the size of this database, we did not retain all entries and all aliases for each entry. Instead, we kept all entries concerning France and all entries corresponding to villages, towns and administrative regions in other coutries, provided their population is known to `GeoNames` and equals at least 200 people. Moreover, we discarded all location names with a non-empty language indication different than "French" or that contained non-French characters. For each retained location name, we store the `GeoNames` id, the `GeoNames` normalized name, and the latitude and longitude. We also compute a weight from the number of inhabitants when it is provided by `GeoNames`. This weight will be used during the NE disambiguation step. Moreover, this weight allows us to compute a reasonable scale level for use in the final interface when showing the location in *Google Maps*. In case of homonymy, we assign unique normalized forms by appending an index to the original normalized form of each

[2] `https://gforge.inria.fr/projects/lingwb/`, distributed under an LGPL license.

[3] In fact, our system is slightly more complex. In particular, it knows that the name of a capital city can be the author of a quotation, but should not be typed *Person* anyway. This is one of the most frequent cases of metonymy, described in introduction.

[4] Freely available at `http://www.geonames.org`

[5] A full dump can be downloaded freely at `http://download.wikimedia.org/frwiki/latest/frwiki-latest-pages-articles.xml.bz2`.

entry but the first one (e.g., there are 14 entries for locations named *Paris*, whose normalized forms are respectively *Paris*, *Paris (2)*,. . . *Paris (14)*).

For other kinds of NEs (persons, organizations, companies, products and brands, artworks), we extracted information from the (French) Wikipedia. The exploitation of Wikipedia for NE detection and resolution is not new, but has been proved efficient [Balasuriya et al., 2009]. We manually defined a mapping from a set of Wikipedia "categories" to one of the above-mentioned NE types. This allowed to type the title of each relevant Wikipedia article. Each typed article gives birth to an entity, whose normalized form is built from the title of the article, to which a disambiguation index may be appended, as for the case of locations. Apart from the title of the article, we extract other mention "variants" by two different means:

- we parse the first sentence of each article and automatically extract variants from it (e.g., *CIA* in addition to *Central Intelligence Agency*, or *Marie-Ségolène Royal* in addition to *Ségolène Royal*); we also extract a "definition" for the entity (in the case of Ségolène Royal, *femme politique française (22 septembre 1953, Dakar –)*).
- we retrieve all redirection pages in the Wikipedia and consider their titles as variants for the entity denoted by the target page.

In the case of person names, additional variants are computed. Indeed, the set of already gathered variants and a large-coverage lexicon of first names extracted from our general-purpose French lexicon allow us to segment person names into the first name, a possible middle name, the last name, and a gender if possible [6]. New variants are then computed, in particular the omission or abbreviation of first and middle names, as in *M.-S. Royal* or *Royal*.[7] As for locations, we assign a weight to all entities extracted from Wikipedia, that will be used during the NE disambiguation step. We compute this weight in a very simple way, based on the size (number of lines) of the whole Wikipedia article.

The output of this extraction process from `GeoNames` and Wikipedia is corrected and enriched by a blacklist and a whitelist of NEs, both manually drawn up. From the resulting NE database, we extract a gazetteer from all variants of each NE, associated with its type (and gender for person names when available).[8]

[6] This person name variant extraction heuristics is specific to names from languages such as French that write the given name before the family name, which is the case in most person name occurrences in our corpus. Other orders should be taken into account in further work.

[7] A candidate such as *Royal*, i.e. an isolated last name, is discarded during the disambiguation step unless it refers to an entity mentioned earlier in the same news item in a more extended form, e.g. *Ségolène Royal*

[8] As part of SXPipe, this database is freely available within the SXPipe distribution.

3.2. NE Recognition

A context-free grammar consisting of 130 rules has been developed for defining patterns based on this gazetteer, as well as on specific lists for identifying relevant contexts (e.g., *ville*, *village*, *localité*, i.e., *city*, *village*, *locality*; another example is a large list of first names, a list of possible titles such as *Dr.*, *Mme*, and others). Disambiguation heuristics have been activated,[9] in order to make the amount of ambiguities added by this NE module as low as possible, although not null. Therefore, the output of the NE Recognizer is a DAG (Directed Acyclic Graph) in which each possible NE span and type combination is represented by a distinct path. Note that recognition involves typing. Therefore, in case of NE type ambiguity, this will be treated as NE recognition ambiguity, and solved by the NE disambiguation and normalization step.

3.3. NE Disambiguation and Resolution

The aim of the NE disambiguation and normalization module is to chose at most one NE reading for each sequence of tokens recognized in the previous step. A *reading* is defined as the combination of a path in the input DAG and an entry in the NE database (i.e., one path in the input DAG may correspond to several readings). Unlike [Pilz and Paaß, 2009] and others, our disambiguation module relies on heuristics based on quantitative and qualitative information, but not on machine learning techniques.

First, we define the *salience level* of an entity as follows. Within a given document, each mention of an entity increments its salience level by its weight. Moreover, we define for each document a geographic context (country and city) by storing all countries and cities mentioned in the document. Each mention of a location entity increments its salience level by an additional (small) weight if it is consistent with the geographic context. On the other hand, we divide by 2 the salience level of all entities each time we move from a document to the next one (here, documents are news wire items).[10]

The strategy we use to select a unique reading, or a limited number of competing readings, can be summed up as follows. NE readings may be tagged as "dubious" or "normal", according to various heuristics that involve the type, the salience level and the surface form of the mention of the entity, as well as its left context. For example, a person name is considered dubious if it is a single token and if it is not the last name of an entity that has been previously detected. If dubious and normal readings share at least one token, dubious readings are discarded.

Among remaining NE readings, we reduce the ambiguity as follows. Among all readings corresponding to the

[9] E.g., longest match heuristics in some cases, preferences between pattern-based and gazetteer-based detection, and others.

[10] This division by 2 is merely a rule-of-thumb choice. In further work, we intend to conduct a series of experiments in order to optimize the salience dynamics w.r.t. performance levels.

Person	Total	Known	Unknown
References	223	111	112
Mentions	672	252	172
Location	Total	Known	Unknown
References	261	217	44
Mentions	672	613	59
Organization	Total	Known	Unknown
References	196	101	95
Mentions	463	316	147

Table 2: Corpus figures

same path, we keep only the reading that corresponds to the entity with the highest salience level.

Finally, in order to retain only one reading, we apply a simple longest-match left-to-right heuristics. However, ambiguity can be preserved in some cases, e.g., if a manual validation step follows the automatic processing.

4. Creation of a Reference Corpus for NER in French

In order to evaluate our NER and resolution system, a subset of news made available to us by the AFP has been selected and manually annotated. This annotation has the form of inline XML tagging and includes both mention-level and reference-level features: span boundaries and type on the one hand, and unique identifier matching a record in the reference base on the other. We aim indeed at addressing NER and reference resolution in an integrated way, which reasonably entails a unique evaluation resource suitable for evaluating the performance of both modules as well as of their interaction (in a comparable way to what is described in [Möller et al., 2004]).

This set of news wires is made up of 100 items with an average of 300 words each. Table 2 shows the distribution over NE types and mentions with known and unknwon refrences.It is freely available within the SXPipe distribution.

The NE types which we included in the annotation are similar to the ones selected in most dedicated conference shared tasks; they are so far limited to *Person*, *Organization* and *Location*. The identifier is a copy of the record identifier in the NER resources (section 3.3.). If no matching record exists, the canonical form is then stated and an attribute indicating this absence of record is added to the mention tag. The NE mentions do not include tokens which are not part of the name itself, such as titles for person names (*Dr.* or *Mr.*).

Annotation examples:

```
- Le président <Person name="Barack
  Obama">Barack Obama</Person> a
  approuvé un accord

- grippe porcine au <Location
  name="Canada (2)">Canada</Location>
  a été révisé

- <Person name="Mohammed Abdullah
  Warsame" ref="unknown">Mohammed
  Abdullah Warsame</Person>, 35 ans,
  habitant
```

5. Evaluation

Development and Test Data The gold corpus was divided in two parts of equivalent size: a development set and a test set. This division between development and test data ensures that the accuracy of the system is not artificially improved by the knowledge of the whole data. One out of every two items forms the test set in order to avoid systematic biaises caused by specific NEs which occur more frequently in a specific part of the corpus.

Metrics As for the metrics applied, we rely on a classical F-score obtained by the harmonic mean of precision and recall. However we calculated more than one F-score, by considering different ways of defining the intersection between the number of entities retrieved by the system and the number of relevant entities occuring in the evaluation corpus.

At the level of NER, it seems indeed reasonable to consider as fully correct any detection of entity mention along with its exact boundaries (correctness of span) and type. This is for example the requirement of the Conll 2003 shared task [Sang and Meulder, 2003] and the scoring is then based on it. Our choice thus depart from scoring methods such as the one of the Message Understanding Conference (MUC) framework [Grishman and Sundheim, 1996] which also allows for partial credit in cases of partial span or wrong type detection. One could also consider that, depending on the application context, one feature or another is of more importance. For instance, extracting entities with incorrect spans or failing to get a maximal precision generates a noise which can be highly unconvenient with regards to the user experience and expectations. Precision can thus be favoured in the calculation of the F-score by giving it more weight than to recall. This focuses efforts on the improvement of precision, even if it means a drop of recall performance, if this is the result considered as the best suited for an application. The metrics used in MUC include this kind of considerations and produce several measures depending on the system feature which is put forward.

When including the performance of reference resolution within the evaluation, the retrieved-relevant intersection must consider as correct matches only the ones which show the right reference as one of their feature. Fully correct matches at the reference resolution level are thus the ones which show span, type and reference correction.

In practice we aim at obtaining three evaluation levels. First, at the recognition level, the accuracy of the system is scored according to the mentions whose span and type are correctly detected. Then we measure the ability of the system to detect that a mention should be matched against a reference record (as opposed to mentions that do not correspond to an entity present in the database). Last, we score

Sub-task	Prec.	Rec.	F-sc.
Detection (span & type are correct?)	0.81	0.77	0.79
Reference detection (entity known?) (among correct span & type)	0.97	0.99	0.98
Reference resolution (which entity?) (among correct span & type & known)	0.91	–	–

Table 3: Evaluation results for the three following subtasks on French news wires: NE recognition, NE reference detection, NE reference resolution.

the accuracy of the resolution step. The set of cases considered for this last measure is therefore the intersection of correctly recognized entities mentions with entities mentions linked to a reference.

Results Table 3 shows the results for the three evaluation levels of our system running on the test gold set. As can be seen, the step that exhibits the lowest F-score is the detection step. The detailed F-scores for Person names, Location names and Organization names are respectively 0.80, 0.85 and 0.68. This can be compared to results of [Brun et al., 2009a], which respectively reach 0.79, 0.76 and 0.65 [Jacquet, p.c.]. The two other steps are difficult to compare to other work, especially for French. However, during the development of NP, we have seen how sensitive the results are to the quality of the NP reference database extraction, as well as to the heuristics used during the resolution step. Therefore, we consider that there is still room for significant improvements both at the detection and at the resolution steps. This includes the conjunction of our rule- and resource-based techniques with machine-learning methods.

6. Future Work

A more finalized reference base, structured by an underlying ontology, shall be used in further versions of NP. We also intend to improve NP's ability to detect new potential entities references appearing in texts. The types of entities handled by the system should be extended beyond the limitations stated in 4. by integrating more entity types, in an extent comparable to the categories defined for the Conll 2003 shared task, in particuliar the *Miscellaneous* category. This integration involves carrying out more annotation and double validation in order for the system to benefit from a complete evaluation resource. More generally, the multilingual production of the AFP should give rise to the development of resources and system adaptation for other languages than French.

Acknowledgments

This work was supported in part by the Scribo project funded by the French "Pôle de compétitivité" System@tic ("Fonds Unique Interministériel") and by the EDyLex project funded by the French National Research Agency (grant number ANR-09-CORD-008).

7. References

ACE08, 2008. Automatic content extraction 2008 evaluation plan.

Balasuriya, Dominic, Nicky Ringland, Joel Nothman, Tara Murphy, and James 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*. Morristown, NJ, USA: Association for Computational Linguistics.

Blume, Matthias, 2005. Automatic entity disambiguation: Benefits to ner, relation extraction, link analysis, and inference. *International Conference on Intelligence Analysis*.

Brun, C., N. Dessaigne, M. Ehrmann, B. Gaillard, S. Guillemin-Lanne, G Jacquet, A. Kaplan, M. Kucharski, C. Martineau, A. Migeotte, T. Nakamura, and S. Voyatzi, 2009a. Une expérience de fusion pour l'annotation d'entités nommées. In *Proceedings of TALN2009*.

Brun, Caroline, Maud Ehrmann, and Guillaume Jacquet, 2009b. A hybrid system for named entity metonymy resolution. In *LNCS 5603, Selected papers from the 3rd Language and Technology Conference (LTC 2007)*. Poznan, Poland: Springer-Verlag.

Grishman, Ralph and Beth Sundheim, 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics (CoLing'96)*. Copenhagen, Denmark.

Murgatroyd, David, 2008. Some linguistic considerations of entity resolution and retrieval. In *Proceedings of LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*.

Möller, Knud, Alexander Schutz, and Stefan Decker, 2004. Towards an integrated corpus for the evaluation of named entity recognition and object consolidation. In *Proceedings of the SemAnnot Workshop at ISWC2004*.

Pilz, Anja and Gerhard Paaß, 2009. Named entity resolution using automatically extracted semantic information. In *Proceedings of LWA 2009*.

Poibeau, Thierry, 2005. Sur le statut référentiel des entités nommées. *CoRR*, abs/cs/0510020.

Sagot, Benoît and Pierre Boullier, 2005. From raw corpus to word lattices: robust pre-parsing processing with SxPipe. *Archives of Control Sciences, special issue on Language and Technology*, 15(4):653–662.

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

Sang, Erik F. Tjong Kim and Fien De Meulder, 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*.