HAL open science

Enriching a Text by Semantic Disambiguation for Information Extraction

Bernard Jacquemin, Caroline Brun, Claude Roux

To cite this version:

Bernard Jacquemin, Caroline Brun, Claude Roux. Enriching a Text by Semantic Disambiguation for Information Extraction. Proceeding of the Workshop on Using Semantics for Information Retrieval and Filtering: State of the Art and Future Research (LREC 2002), May 2002, Las Palmas, Canary Islands, Spain. pp.45-51. halshs-00003968

HAL Id: halshs-00003968

https://shs.hal.science/halshs-00003968v1

Submitted on 20 Jun 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Enriching a Text by Semantic Disambiguation for Information Extraction

B. Jacquemin – C. Brun – C. Roux
Xerox Research Centre Europe
6, chemin de Maupertuis
38 240 Meylan
ILPGA – Sorbonne Nouvelle
19, rue des Bernardins
75 005 Paris
{Jacquemin,Brun,Roux}@xrce.xerox.com

halshs-00003968, version 1 - 20 Jun 2005

Abstract

External linguistic resources have been used for a very long time in information extraction. These methods enrich a document with data that are semantically equivalent, in order to improve recall. For instance, some of these methods use synonym dictionaries. These dictionaries enrich a sentence with words that have a similar meaning. However, these methods present some serious drawbacks, since words are usually synonyms only in restricted contexts. The method we propose here consists of using word sense disambiguation rules (WSD) to restrict the selection of synonyms to only these that match a specific syntactico-semantic context. We show how WSD rules are built and how information extraction techniques can benefit from the application of these rules.

Keywords: question answering (QA) - information extraction (IE) - information retrieval (IR) - word sense disambiguation (WSD) - electronic dictionary - semantic dictionaries - synonymy - morpho-syntactic derivation - paraphrase - query expansion - ontologies - lexical resources.

1 Introduction

In today's world, the society of communications is gaining in importance every day. The amount of electronic documents – mainly by Internet, but not only – grows more and more. With this increase, no one is able to read, classify and structure those documents so that the requested information can be reached when it is needed. Therefore we need tools that reach a shallow understanding of the content of these texts to help us to select the requested data.

The process of understanding a document consists in identifying the concepts of the document that correspond to requested information. This operation can be performed with linguistic methods that permit the extraction of various components related to the data that are requested.

Since the beginning of the '90s, several research projects in information extraction from electronic text have been using linguistic tools and resources to identify relevant elements for a request. The first ones, based on domain-specific extraction patterns, use hand-crafted pattern dictionaries (CIRCUS [Lehnert, 1990]). But systems were quickly designed to build extraction pattern dictionaries automatically. Among these systems, AutoSlog [Riloff, 1993, Riloff and Lorenzen, 1999] builds extraction pattern dictionaries for CIRCUS. CRYSTAL [Soderland et al., 1995] creates extraction patterns lists for BADGER, the successor of CIRCUS. These learners use hand-tagged specific corpora to identify structures containing the relevant information. The syntactic structure used by CRYSTAL is more subtle than the one used by AutoSlog. CRYSTAL is able to make the most of semantic classes. WHISK [Soderland, 1999] is one of the most recent information extraction system. WHISK has been designed to learn which data to extract from structured, semi-structured and free text [1]. A parser and a semantic tagger have been

[1] We use the term "structured text" to refer to what the database community calls semi-structured text; "semi-structured text" is ungrammatical and often telegraphic text

implemented for free text. This system is the only one to process all of these three categories of text.

These methodologies need domain-specific pattern dictionaries that must be built for each different kind of information. However, none of these methods can be directly applied to generic information. Thus we decide to bypass these two obstacles: our approach is based on the utilization of an existing electronic dictionary, in order to expand the data in a document to equivalent forms extracted from that dictionary.

Our method deals with the identification of semantic contents in documents through a lexical, syntactic and semantic analysis. It then becomes possible to enrich words and multi-word expressions in a document with synonyms, synonymous expressions, semantic information etc. extracted from the dictionary.

2 Problems and Prospects

As for a lot of methodologies developed for natural language processing, the results of a method of information extraction are evaluated by two measures: precision and recall. Precision is the ratio of correctly extracted items to the number of items both correctly and erroneously extracted from the text; noise is the ratio of the faulty extracted items to all the achieved extractions. Recall is the ratio of correctly extracted items to the number of items actually present in the text. The problem consists in improving both precision and recall.

2.1 Recall improvement

A usual technique to improve the recall consists of enriching a text with a list of synonyms or near-synonyms for each word of that text. For example, all the synonyms of "climb" would be added to the document, even though some of those meanings have a remote semantic connection to the text. By this kind of enrichment, all the ways to express the same token (but not the same meaning) are taken into account.

This type of enrichment can be extended to synonymous expressions with a robust parser: syntactic dependencies and their arguments (the tokens belonging to the selected expression) are enlarged to dependencies that are generated out of the corresponding synonymous expressions.

The recall is usually optimised to the detriment of the precision with those techniques, since most words within a set of synonyms are themselves polysemous and are seldom equivalent for each of their meanings. Thus, a simply adding of all those polysemous synonyms in a document introduces meaning inconsistencies. Noise may stem from these inconsistencies.

2.2 Reduction of noise – Precision improvement

We notice that improving the recall using synonyms may often increase the noise. Although identified in the domain of IE, this problem is not yet solved and it has a negative influence on the system effectiveness. Our purpose is to use the linguistic context of the polysemous tokens to identify their meanings and select contextual synonyms or synonymous expressions. This approach should improve the precision in comparison with adding all the synonyms.

Sentences in the text:

La température **grimpe**.
(The temperature is climbing.)

Corresponding set of synonyms:

escalader	*monter*
(to climb)	*(to go up)*
sauter	*augmenter*
(to jump)	*(to increase)*
se hisser sur	
(to heave oneself up onto)	

Sentences resulting from the enrichment:

La temperature escalade.
La temperature ***monte****.*
La temperature saute.
La temperature ***augmente****.*
La temperature se hisse sur (???).

Figure 1: Enrichment by a list of synonyms.

that does not follow any rigid format; "free text" is simply grammatical text [Soderland, 1999].

For example, the dictionary [2] entry for the word *grimper* contains a set of 5 synonyms. If we use these synonyms to enrich the original text, we obtain five variations of the original sentence. Only the second and the fourth of the enriching variations are accurate in this context. The meteorological context associated with the word *température* in the dictionary should correctly discriminate the synonyms in this context: in the dictionary, each synonym of a lemma is associated with a meaning of this lemma and with the typical linguistic context of the lemma in this sense.

Consequently, we decided to use the linguistic context of the words that can be enriched to discriminate which synonyms should be used and which should not. The synonyms are stored in the dictionary according to the sense of each lemma. So, the task amounts to performing a lexical semantic disambiguation of the text and using synonymous expressions in the selected meanings to enrich the document.

3 Enrichment method by WSD

3.1 Our experience in WSD

We previously have developed a range of tools and techniques to perform Word Sense Disambiguation (WSD), for French and English. The basic idea is to use a dictionary as a tagged corpus in order to extract semantic disambiguation rules, [Brun et al., 2001, Brun, 2000, Brun and Segond, 2001, Dini et al., 1998, Dini et al., 2000]. Since electronic dictionaries exist for many languages and they encode fine-grained reliable sense distinctions, be they monolingual or bilingual, we decided to take advantage of this detailed information in order to extract a semantic disambiguation rule database [3]. The disambiguation rules associate each word with a sense number taking the context into account. For bilingual dictionaries the sense number is associated with a translation, for monolingual dictionaries with a definition. WSD is therefore performed according to sense distinctions of a given dictionary. The linguistic rules have been created using functional dependencies provided by an incremental shallow parser (IFSP, [Aït-Mokhtar and Chanod, 1997]), semantic tags from an ontology (45 classes from WordNet [Fellbaum, 1998] for English) as well as information encoded in SGML tags of dictionaries. This method comprises two stages, rule extraction and rule application.

- Rule extraction process: for each entry of the dictionary, and then for each sense of the entry, examples are parsed with the IFSP shallow parser. The shallow parsing task includes tokenization, morphological analysis, tagging, chunking, extraction of functional dependencies, such as subject and object (SUBJ(X, Y), DOBJ (X, Y)), etc. For instance, parsing the dictionary example attached to one particular sense $\mathbf{S}_i$ of *drift* :

  1)The country is drifting towards recession.

  Gives as output the following chunks and dependencies :

  [SC [NP The country NP]/SUBJ :v is drifting SC] [PP towards recession PP] SUBJ(country, drift) VMODOBJ(drift, towards, recession)

  Using both the output of the shallow parser and the sense numbering from the dictionary we extract the following semantic disambiguation rule: When the ambiguous word "drift" has *country* as subject and/or *toward recession* as modifier, it can be disambiguated with its sense $\mathbf{S}_i$. We repeat this process as all dictionary example phrases in order to extract the word level rules, so called because they match the lexical context.

  Finally, for each rule already built, we use semantic classes from an ontology in order to generalize the scope of the rules. In the above example the subject "country" is replaced in the semantic disambiguation rule by its ambiguity class. We call ambiguity class of a word, the set of WordNet tags associated with it. Each word level rule generates an associated class level rule, so called because it matches the semantic context: when the ambiguous

[2] The dictionary we use is a French electronic one [Dubois and Dubois-Charlier, 1997]. We will give a more detailed information about it later.

[3] The English dictionary contained 39 755 entries and 74 858 senses, ie a polysemy of 1.88; the French dictionary contained 38 944 entries and 69 432 senses, ie a polysemy of 1.78

word "drift" has a word belonging to the WordNet ambiguity class *noun.location* and *noun.group* as subject and/or a word belonging to the WordNet ambiguity class *noun.shape*, *noun.act*, and *noun.state* as modifier, it disambiguates with its sense $\mathbf{S}_i$. Once all entries are processed, we can use the disambiguation rule database to disambiguates new unseen texts. For French, semantic classes (69 distinctive characteristics) provided by the *AlethDic* dictionary [Gsi-Erli, 1993] have been used with the same methodology.

- Rule application process: The rule applier matches rules of the semantic database against new unseen input text using a preference strategy in order to disambiguate words on the fly. Suppose we want to disambiguate the word drift, in the sentence:

  2) In November 1938, after Kristallnacht, the world drifted towards military conflict.

  The dependencies extracted by the shallow parser, which might lead to a disambiguation, i.e., which involve *drift*, are:

  SUBJ(world, drift)
  VMODOBJ(drift, towards, conflict)

  The next step tries to match these dependencies with one or more rules in the semantic disambiguation database. First, the system tries to match lexical rules, which are more precise. If there is no match, then the system tries the semantic rules, using a distance calculus between rules and semantic context of the word in the text [4]. In this particular case, the two rules previously extracted match the semantic context of *drift*, because *world* and *country* shares semantic classes according to WordNet, as well as *conflict* and *recession*.

The methodology attempts to avoid the data acquisition bottleneck observed in WSD techniques. Thanks to this methodology, we built all-words (within the limits of the used dictionary) unsupervised Word Sense Disambiguator for French (precision: 65%, recall: 35%) and English (precision: 79%, recall: 34%).

[4] The first parameter of this metric is the intersection of the rule classes and the context classes; the second one is the union of the rule classes and the context classes. Distance equals the ratio of intersection to union.

3.2 Xerox Incremental Parser (XIP)

IFSP, which was used in the first experiments on semantic disambiguation at Xerox, has been implemented with transducers. Transducers proved to be an interesting formalism to implement quickly an efficient dependency parser, as long as syntactic rules would only be based on POS. The difficulty of using more refined information, such as syntactic features, drove us to implement a specific platform that would keep the same strategies of parsing as in IFSP, but would no longer rely on transducers.

This new platform [Aït-Mokhtar et al., 2001, Roux, 1999] comprises different sorts of rules that chunk and extract dependencies from a sequence of linguistics tokens, which is usually but not necessarily a sentence. The grammar of French that has been developed computes a large number of dependencies such as *Subject, Object, Oblique, NN* etc. These dependencies are used in specific rules, the disambiguation rules, to detect the syntactic and semantic information surrounding a given word in order to yield a list of words that are synonyms according to that context. Thus, a disambiguation rule manipulates together a list of semantic features originating from dictionaries, and a list of dependencies that have been computed so far. The result is a list of contextual synonyms.

If (Dependency$_0$(t, t^0) & ... & Dependency$_n$(t,t^k) & ... attribute$_p$(t^j)=v^u)

synonym(t) = s^0,...,s^n.

where

t^0,...,t^n is a list of token

s^0,...,s^n a list of synonyms.

The contextual synonymy between *grimper* and *augmenter* can be defined with the following rule. The feature *MTO* is one of the semantic features that are associated with the entries of the Dubois dictionary. This feature is associated with each word that is connected to meteorology, such as *chaleur, froid, température* (heat, cold, température).

if (Subject(*grimper*, X) AND feature(X, *domain*)=MTO) synonym(*grimper*) = *augmenter*.

This rule applies on the above first example, *La température grimpe*, but fails to apply on the third sentence, *L'alpiniste grimpe le mont Ventoux*, since the subject does not bear the MTO feature.

Example:

- La température grimpe.
  (the temperature is climbing)
- La température augmente.
  (the temperature is rising)
- L'alpiniste grimpe le mont Ventoux.
  (the alpinist climbs the mount Ventoux)
- ???L'alpiniste augmente le mont Ventoux.
  (???the alpinist raises the mount Ventoux)

Figure 2: Application of a disambiguation rule for enrichment.

3.3 Which WSD for which enrichment?

3.3.1 A very rich dictionary information

The new robust parser offers a flexible formalism and the possibility to handle semantic or other features. In addition to this parser, the semantic disambiguation now uses a monolingual French dictionary [Dubois and Dubois-Charlier, 1997]. This dictionary contains many kind of information in the lexical field as well as in the syntactic or the semantic one. From the 115 229 entries of this dictionary, we can only use the 38 965 ones that are covered by the morphological analyser. These entries represent 68 588 senses, ie a polysemy of 1.76.

We build lexico-syntactic WSD rules using the methodology presented above (cf. section 3.1): examples of the dictionary are parsed; extracted syntactic relations and their arguments are used to create the rules. We also make the most of the domain indication (171 different domains) to generalize the example rules (see later for details) – as previously done using WordNet for the English WSD and by AlethDic for the French one [Brun et al., 2001].

We use the specificity of the dictionary to improve the disambiguation task as far as possible in order to maximize the enrichment of the documents. The information of this dictionary is divided into several fields: domain, example, morphological variations, derived or root words, synonyms, POS, meaning, estimate of use frequency in the common language; in the verbal part of the dictionary only, syntactico-semantic class and subcategorization patterns of the arguments of the verb. Resulting WSD rules are spread over three levels reflecting the abstraction register of the dictionary fields.

3.3.2 Disambiguation rules at various levels

We build a disambiguation rule database at three levels: rules at word level (23 986), rules at domain level (22 790) and rules at syntactico-semantic level (40 736).

Word level rules use lexical information from the examples. They correspond to the basic rules in the previous system, which use constraints on words and syntactic relations. These dependencies are extracted from the illustrative examples from the dictionary.

L'avion de la société **décrit** un large cercle avant de (...)
*(The company's plane **describes** a wide circle before (...))*
SUBJECT(décrire,avion)
OBJECT(décrire,cercle)

Example in the dictionary for the entry "décrire":
L'avion décrit un cercle.
(The plane describes a circle.)
SUBJECT(décrire,avion)
OBJECT(décrire,cercle)

Figure 3: WSD at word level.

Rules at domain level are generalized from word level rules: instead of using the words of the examples as arguments of the syntactic relations in the rules, we replace them by the domains they belong to. These rules correspond to the class level rules in the previous system, but an improvement in comparison with them is that in some cases, we can discriminate the right domain if the argument is polysemous. This is mainly due to the internal consistency of the dictionary that enables the correspondences of domain across different arguments of a dependency. The consistency should help to reduce the noise.

We don't rule out the possibility of using other lexico-semantic resources to generalize or expand

L'escadrille décrit son approche vers l'aéroport où (...)
(The squadron describes its approach to the airport where (...))
SUBJECT(décrire,escadrille[dom:AER])
OBJECT(décrire,approche[dom:LOC])

Example in the dictionary for the entry "décrire":
L'avion décrit un cercle.
(The plane describes a circle.)
SUBJECT(décrire,avion[dom:AER])
OBJECT(décrire,cercle[dom:LOC])

Figure 4: WSD at domain level.

L'escadrille décrit son approche vers l'aéroport where (...)
(The squadron describes its approach to the airport where(...))
SUBJECT(décrire,escadrille[dom:AER])
OBJECT(décrire,approche[dom:LOC])

Subcategorisation for the entry "décrire":
Transitive verb;
Subject inanimate.
SUBJECT(décrire,?[subcat:inanimate]) & OBJECT(décrire,?)

Figure 5: WSD at lexico-semantic level.

this kind of rules, as we did previously using French EuroWordNet or AlethDic. These lexicons present the advantage of a hierarchical structure that doesn't exist for the domain field in the Dubois dictionary. Nevertheless, we will encounter the problem of the mapping of the various resources used by the system to avoid inconsistencies between them, as shown in [Ide and Véronis, 1990, Lux et al., 1999, Brun et al., 2001].

The third level of the rules currently in use in the semantic disambiguator is the syntactico-semantic one. The abstraction level of these rules is even higher than in the domain level. They are built from a syntactic pattern of subcategorization that indicates the typical syntactic construction of the current entry in its current meaning. Although the distinction between the arguments is very general – they are differentiated from human, animal and inanimate – our examination of the verbal dictionary indicates that, for 30% of the polysemous entries, this kind of rules is sufficient to choose the appropriate meaning.

3.4 Enrichment at various levels

WSD is not an end in itself. In our system, it is a means to select appropriate information in the dictionary to enrich a document. The quality and the variety of this enrichment vary according to the quality and the richness of the information in the dictionary. The variety of information allows several kind of enrichment.

For the specific task of information extraction, an index of the documents whose information is likely to be extracted is built. It allows the classification of all the linguistic realities extracted from text analysis. These realities are listed according to the XIP-formalism: syntactic relations, arguments, and features attached to the arguments. The enrichment is done inside the index because dependencies can be added without affecting the original document.

3.4.1 Lexical level

Replacing a word by its contextual synonyms is the easiest way to perform enrichment. This method of recall improvement is very common in IE, but in our system, the enrichment is targeted according to the context thanks to the semantic disambiguation. This process often reduces the noise. The enrichment is achieved by copying the dependencies containing the disambiguated word and by replacing this word by one of its synonyms.

3.4.2 Lexico-syntactic level

The lexico-syntactic level of enrichment is more complex to achieve. The task consists in replacing a word by a multi-word expression (more than 14 000 synonyms are multi-word expressions in our dictionary) or in replacing a multi-word expression by a word, taking into account the words (lexical) and the dependencies between them (syntactic):

- Replacing a word by a multi-word expression (see figure 7):

La température grimpe.
(The temperature is climbing.)

Original index:
SUBJECT(grimper,température)

Set of targeted synonyms:
monter, augmenter.

Enriched index:
SUBJECT(grimper,température)
SUBJECT(monter,température)
SUBJECT(augmenter,température)

Figure 6: Enrichment at lexical level.

- Parse the multi-word expression to obtain dependencies;
- Match the corresponding dependencies in the text;
- Instantiate the missing arguments with the text arguments.

- Replacing a multi-word expression by a word:
  - Identify the POS of the word;
  - Select dependencies implying one and only one word of the multi-word expression;
  - Eliminate dependency where this word has a different POS;
  - Replace this word with its synonym in the remaining dependencies.

Since our work is based on the Dubois dictionary – whose entries are single words – most of the enrichment is one-to-one word. When a multi-word expression appears in the synonyms list, a single word has to be replaced by a multi-word expression, and the inverse process can be achieved if necessary. The complex case of replacing a multi-word expression by another multi-word expression could arise, but we never encounter this situation. The replacement of a multi-word expression by another is not yet implemented because of the complexity of the process. Nevertheless, the system relies on relations and arguments that are easy to handle, very simple and modular. These characteristics should allow us to bypass the inherent complexity of these structures.

Le spécialiste a édité un manuscrit très abîmé.
(The specialist published a very damaged manuscript.)

Original index:
SUBJECT(éditer,spécialiste)
OBJECT(éditer,manuscrit)

Targeted synonymous expression:
établir l'édition critique de

Extracted dependencies from the expression:
SUBJECT(établir,?)
OBJECT(établir,édition)
EPITHET(édition,critique)
PP(édition,de,?)

Enriched index:
SUBJECT(éditer,spécialiste)
OBJECT(éditer,manuscrit)
SUBJECT(établir,spécialiste)
OBJECT(établir,édition)
EPITHET(édition,critique)
PP(édition,de,manuscrit)

Figure 7: Enrichment at lexico-syntactic level.

3.4.3 A semantic level example

Syntactico-semantic fields in the dictionary allow a third enrichment level. The syntactico-semantic class structure contains very useful information that makes it possible to link verbs that are semantically related but lexically and syntactically very different. It might be interesting to semantically link *vendre* ("to sell", class D2a) and *acheter* ("to buy", class D2c) even though their respective actors are inverted. For example, *le marchand vend un produit au client* (the trader sells a product to the customer) bears the same meaning as *le client achète un produit au marchand* (the customer buys a product from the trader). The semantic class gives a general meaning of the verb(D2, meaning *donner, obtenir*, to give, to obtain), while the syntactic pattern (a for *vendre*: *fournir qc à qn*, to sup-

ply so with sth, transitive with a *à* oblique compliment, c for *acheter*: *prendre qc à qn*, to take sth to so, transitive with a *à* oblique compliment) yields the semantic realization.

Le papa offre un cadeau à sa fille.
(The father is giving a present to his daughter.)

Original index:
SUBJECT(offrir,papa)
OBJECT(offrir,cadeau)
OBLIQUE(offrir,fille)

offrir 01: D2a (to give sth to sb)
D2a corresponds to D2e (receive, obtain sth from sb).
recevoir 01: D2e

Enriched index:
SUBJECT(offrir,papa)
OBJECT(offrir,cadeau)
OBLIQUE(offrir,fille)
SUBJECT(recevoir,fille)
OBJECT(recevoir,cadeau)
????(recevoir,de,papa)

Figure 8: Enrichment at semantic level.

In a same perspective, a syntactico-semantic class constitutes another synonym set. Since this set is too general and too imprecise, it cannot be used to enrich a document. Still, it can be used as a last resort to enrich the query side when other methods have failed. We will not use this set as enrichment, but only to match a query by the class if the enrichment fails.

4 Evaluation

Though the method presented in this article is based on previous works, the use of other tools and lexical resource may have extended the potential of WSD rules. In particular, it is possible that the number of domains increase precision, and the use of subcategorization patterns may ensure more general rules to increase recall.

The partial evaluation we performed concerns 604 disambiguations in a corpus of 82 sentences from the French newspaper *Le Monde*. Precision in WSD is ratio of correct disambiguations to all disambiguations performed; recall is ratio of correct disambiguations to all possible disambiguations in the corpus. We distinguish the mistakes due to the method and the ones linked to our analysis tools in order to identify what we have to improve in order to increase the performance. These results are promising since both precision and recall are better than in the previous system.

Tokenization mistakes	44	7.28%
Tagging mistakes	19	3.15%
Parsing mistakes	9	1.49%
WSD mistakes	84	13.91%
Precision	**448**	**74.17%**
Recall		**43.61%**

Table 1: WSD method evaluation.

We note some remarks about this evaluation:

1. The lexicon used to perform tokenization has been modified in order to include additional information from the dictionary. We noticed during this evaluation some problems of coverage;

2. For this first prototype, we do not yet establish a strategy for cases in which multiple rules match. If more than one rule can be applied to the context, the sense is randomly chosen among the ones suggested by the matching rules [5];

3. Conversely, we do not yet try a strategy using the domain of disambiguated words as a general context to choose the corresponding meaning of a word to disambigate.

During the evaluation, we also notice that when a result was correct, the suggested synonymous expressions were always correct for the disambiguated word in this context. Our method for an optimized enrichment is validated.

[5] This random choice is only performed for this evaluation and not in a IE perspective, since noise is better than silence in this field.

5 Conclusion

In this paper, we present an original method for processing documents, preparing the text for information extraction. The goal of this processing is to expand each concept by the largest list of contextualy synonymous expressions in order to match a request corresponding to this concept.

Therefore, we implement an enrichment methodology applied to words and multi-word expressions. In order to perform the enrichment task, we have decided to use WSD to contextually identify the appropriate meaning of the expressions to expand. Inconsistent enrichment by synonyms is currently known as a major cause of noise in Information Extraction systems. Our strategy lets the system target the enriching synonymous expressions according to the semantic context. Moreover, this enrichment is achieved not only with single synonymous words, but also with multi-word expressions that might be more complex than simple synonyms.

The WSD task and the resulting enrichment stage are achieved using syntactic dependencies extracted by a robust parser: the WSD is performed using lexico-semantic rules that indicate the preferred meaning according to the context. The linguistic information extracted from the analysis of the documents is indexed for the IE task. This index also stores additional new dependencies stemming from the enrichment process.

The utilization of a unique, all-purpose dictionary to achieve WSD and enrichment ensures the consistency of the methodology. Nevertheless, the information quality and richness of the dictionary might determine the system effectiveness.

The evaluation validates the quality of our method, which allows a great deal of lexical enrichment with less noise than is introduced by other enrichment methods. We have also indicated some ways our method could be expanded and our analysis tools could be improved. Our next step will be to test the effect of the enrichment in an IE task.

The method is designed to achieve a generic IE task, and the tools and resources are developed to process text data at a lexical level as well as at a syntactic or semantic level.

References

[Aït-Mokhtar and Chanod, 1997] Aït-Mokhtar, S. and Chanod, J.-P. (1997). Subject and object dependecy extraction using finite-state transducers. In *ACL'97 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications*, pages 71–77.

[Aït-Mokhtar et al., 2001] Aït-Mokhtar, S., Chanod, J.-P., and Roux, C. (2001). A multi-input dependency parser. In *Proceedings of the Seventh International Workshop on Parsing Technologies*, pages 201–204, Beijing, China. IWPT-2001, Tsinghua University Press.

[Brun, 2000] Brun, C. (2000). A client/server architecture for word sense disambiguation. In *Proceedings of Coling'2000*, pages 132–138, Saarbrücken, Deutschland.

[Brun et al., 2001] Brun, C., Jacquemin, B., and Segond, F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale. *Traitement Automatique des Langues*, 42(3):667–690.

[Brun and Segond, 2001] Brun, C. and Segond, F. (2001). Semantic encoding of electronic documents. *International Journal of Corpus Linguistics*.

[Dini et al., 1998] Dini, L., Di Tomaso, V., and Segond, F. (1998). Error driven word sense disambiguation. In *Proceedings of the Conference COLING-ACL'98*, pages 320–324, Montréal. COLING-ACL.

[Dini et al., 2000] Dini, L., Di Tomaso, V., and Segond, F. (2000). Ginger ii: an example-driven word sense disambiguator. *Computer and the Humanities. Special Issue on SENSEVAL*, 34(1-2):121–126.

[Dubois and Dubois-Charlier, 1997] Dubois, J. and Dubois-Charlier, F. c. (1997). *Dictionnaire des verbes français*. Larousse, Paris. This dictionary exists in an electronic version and is accompanied by the corresponding electronic *Dictionnaire des mots français*.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: an electronic lexical database*, chapter Semantic

Network of English Verbs, pages 69–104. The MIT Press, Cambridge, Massachusetts.

[Gsi-Erli, 1993] Gsi-Erli (1993). *Le dictionnaire AlethDic*. Gsi-Erli, France, 1.5 edition.

[Ide and Véronis, 1990] Ide, N. and Véronis, J. (1990). Mapping dictionaries: A spreading activation approach. In *Proceedings of the 6th Annual Conference of the Centre for the New Oxford English Dictionary*, pages 52–64, Waterloo, Ontario.

[Lehnert, 1990] Lehnert, W. (1990). Symbolic/subsymbolic sentence analysis: Exploiting the best of two worlds. In Barnden, J. and Pollack, J., editors, *Advances in Connexionist and Natural Computation Theory*, volume 1, pages 135–164. Ablex Publishers, Norwood, NJ.

[Lux et al., 1999] Lux, V., Jean, C., and Segond, F. (1999). Wsd evaluation and the looking glass. In *Proceedings of TALN-99*, Cargese. TALN-99.

[Riloff, 1993] Riloff, E. (1993). Automatically constructing a dictionary for information extracting tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816. AAAI Press / MIT Press.

[Riloff and Lorenzen, 1999] Riloff, E. and Lorenzen, J. (1999). Extraction-based text categorization: generating domain-specific role relationships automatically. In Strzalkowski, T., editor, *Natural Language Information Retrieval*. Kluwer Academic Publisher.

[Roux, 1999] Roux, C. (1999). Phrase-driven parser. In *Proceedings of VEXTAL'99*, pages 235–240, Venezia, Italia. VEXTAL'99.

[Soderland, 1999] Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1):233–272.

[Soderland et al., 1995] Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W. (1995). Crystal: Inducing a conceptual dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1314–1320. IJCAI-95.