Evaluating resource acquisition tools for Information Extraction

Thierry Poibeau*, Dominique Dutoit and Sophie Bizouard*****

* THALES RESEARCH AND TECHNOLOGY and LIPN
Domaine de Corbeville, 91 404 Orsay, France
thierry.poibeau@thalesgroup.com

** MEMODATA and CRISCO
17, rue Dumont d'Urville, 14 000 Caen, France
memodata@wanadoo.fr

*** CRIM/INALCO
2, rue de Lille, 75007 Paris, France
sophie.bizouard@club-internet.fr

Abstract

This paper evaluates two different approaches for the elaboration of semantic classes. The framework is an Information Extraction, which needs large amount of domain-dependent resources. An endogenous approach (corpus-based learning) is contrasted with a heterogeneous one (the use of a large semantic network). The two techniques are evaluated.

Cet article vise à évaluer deux approches différentes pour la constitution de classes sémantiques. Nous nous plaçons dans la perspective d'une application d'extraction d'information, pour laquelle la notion de classe sémantique est primordiale. Une approche endogène (acquisition à partir d'un corpus) est contrastée avec une approche exogène (à travers un réseau sémantique riche). L'article présente une évaluation fine de ces deux techniques et leur complémentarité possible.

Keywords – Mots Clés

Semantic classes, Evaluation, Information Extraction, Resources, Semantic network

Classes sémantiques, Évaluation, Extraction d'information, Ressources, Réseau sémantique

1 Introduction: general resources *vs.* specific resources

This paper presents an evaluation methodology applied to the elaboration of semantic classes. Our framework is Information Extraction: in this domain, one needs to develop extraction patterns (on the syntactic level) enriched with semantic classes (on the semantic level). In this introduction we consider the two main approaches that are commonly used to define semantic classes: the access to large general resources and the semi-automatic elaboration form a corpus analysis.

The utility of general resources has been disputed in the last few years. The main objections are enumerated in D. Dutoit's thesis (Dutoit, 2000):

- General linguistic resources (especially large semantic networks) are rarely usable in context since they encode abstract data but not the main linguistic features specific of the domain to be addressed.
- Even if relevant information has been encoded, this information is lost among a large set of irrelevant information. Finding and using the good information is then a difficult task.
- The information that has to be encoded is so large that the task is potentially infinite and, so, is doomed to failure (Victorri, 1998).

The usability of a linguistic resource is largely dependent of the way it is structured. From this point of view, one must notice that most of the above criticisms concerned the Princeton version of Wordnet (Fellbaum, 1998). A large number of people agree on the fact that this semantic network is not sufficiently structured to be optimally used in various operational contexts. On a more theoretical level, message understanding is a process implying different kind of knowledge and the use of a semantic network, even if it is incomplete, lead to a more realistic situation compared to the sole used of the corpus as a reservoir of knowledge. The immediate context is unable to provide the full knowledge implied by an understanding process.

However, several researchers refused these arguments and have investigated some automatic procedure to automatically derived semantic classes from a representative corpus (Hirschmann *et al*, 1975) (Grishman et Sterling, 1994). These experiments are, for the most part, based on a distributionalist assumption (Harris, 1951): in a sublanguage, regular syntactic structures allow to highlight word families occurring in common contexts. These distributional word families are a basis for the elaboration of semantic classes.

This approach is attractive but the results vary a lot wen we take into account the size and the regularity of the corpus on which the acquisition is performed. Papers in this domain rarely address this issue, and most of the experiments are done on very regular corpora (for example Nazarenko *et al.* (2000) report an experiment on medical texts, Faure (2000) on cooking recipes). Moreover, most of semantic classes semi-automatically derived from a corpus are not completely satisfactory (Nazarenko *et al.*, 2001): they cannot be directly injected in an operational application. Several well-known reasons can be given:

- Even in a sublanguage, a syntactic schema can be ambiguous ;
- Words appearing in different contexts are linked by various kinds of relations. These relations can be specific to a given context. The decision to group or not the different words in a class highly depends on the fineness of the expected classification and cannot be automatically *a priori* determined.
- Some words that are very frequent and very general induce semantically irrelevant classes.

It is then necessary to inject some exterior knowledge in the acquisition systems to strengthen the results and develop operational solutions. The learning process is then supervised, like in the experiment from Nazarenko *et al.* (2001) who propose to adapt the medical nomenclature SNOMED to the MENELAS corpus. The authors project the categories of SNOMED on the results of the distributional analysis, which contributes to a refinement of the semantic classes that were otherwise automatically obtained from the corpus. Morin and Jacquemin (1999) use the same kind of technique to specialize a thesaurus (AGROVOC) in relation with a technical corpus.

The debate is then still alive between supporters of general resources versus supporters of a more dynamic approach. We propose to re-investigate this question in this paper. We will take an applicative framework to compare two different approaches for the acquisition of semantic classes. We firstly evaluate the ASIUM system that proposes an automatic acquisition from a corpus and then an interactive validation phase (section 3). We will do the same experiment with the semantic network from MEMODATA that is currently, as far as we know, the richer semantic network for French (section 4). We will conclude on the complementarity of the two approaches (section 5).

2 Experiment protocol

The experiments that are described here have been performed on a financial corpus. 300 news stories reporting company transactions in French have been selected (this constitutes the FIRSTINVEST corpus).

The set of news stories is divided in two unequal parts: 57 stories are reserved for test, 243 for training.

An information engineer, Sophie Bizouard[1], developed an Information extraction system on that domain in Thales. The system included manually elaborated semantic classes that have been considered as a reference for the further experiments we have done in this domain.

3 An interactive symbolic machine learning approach for the automatic acquisition of semantic classes

We have evaluated the contribution of a machine learning system to acquire knowledge from texts and elaborate the resources dedicated to an Information Extraction system. We gave below a brief description of the ASIUM system that we used (for a more detailed description, please refer to (Faure et Nédellec, 1998), (Faure, 2000)).

[1] Sophie Bizouard (INALCO/CRIM) led for the larger part these experiments in Thales Research and Technology in 2001. She ensured a neutral point of view since she did not previously work on Asium nor on The Integral Dictionary.

3.1 The ASIUM acquisition system

The ASIUM system provides knowledge acquired from unstructured texts that have previously been syntactically analyzed. The approach is based on a distributional analysis. ASIUM aims at providing to the analyst a set of classes relevant for the task he is doing and it also provides means to structures these classes into a domain ontology. This ontology gives a clear image of the relationships among classes.

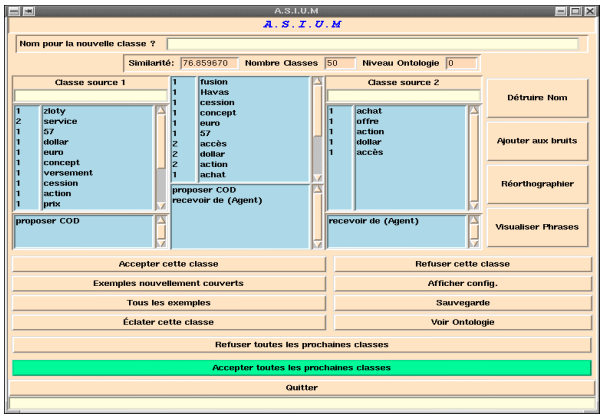

Figure 1. The ASIUM system

The method is original in that it provides an interactive validation phase. The analysis rests on a distributional analysis that allows to generate classes of words appearing in similar contexts (the *base classes*). The similarity measure implemented in ASIUM gives a score for the coverage of a class compared to another one. The system then proposes to group together classes that have a similarity measure above a certain threshold (this threshold is empirically fixed by the expert). The base classes are progressively aggregated following a cooperative bottom-up method.

This classification process is not limited to lists of words appearing in similar contexts but it also extends these lists by induction. Indeed, from two classes C1 and C2 found in two different contexts CTXT1 and CTXT2, one can authorize non-attested sequences like CTXT1 C2 or CTXT2 C1. For example, imagine that *acquisition* and *fusion* belong to the same base class. Then, if *procéder à une acquisition* is a sequence present in the corpus, the system will infer that *procéder à une fusion* is also a valid sequence, even if it is not present in the corpus. The result is then a generalization of the knowledge base that can be directly inferred from the corpus.

ASIUM requires that a domain expert intervene to validate the results. This validation is done by examining each class, one by one, until the threshold is reached. Classes under the threshold are supposed to be non relevant and are automatically rejected by the system.

3.2 From unsupervised to supervised methods

The validation phase is very important in the acquisition process designed in ASIUM. Given that the amount of data was reduced, the threshold that we retained to merge the base classes was low. The consequence is that the propositions from the system concerned classes having very few elements in common.

To solve this problem, we made two different experiments. In the first one, ASIUM is used as it. The propositions of the system are the result of an unsupervised process: they are solely based on the number of common elements between the different base classes. This strategy leads to an expensive validation task: the analyst needs to validate a large number of irrelevant items.

A second experiment then used a new function of ASIUM allowing to focalize the validation process on the sole classes containing words that are relevant for the domain. These words are defined by the analyst in a file that is then considered as a filter. For example, to find the semantic class `purchase_operations` (`Opération d'achat`), the filter contained the following set of words: *achat*, *fusion*, *scission* and for the class `Company` (`Entreprise`), the words: *entreprise*, *société*, *filiale*.

This strategy led us to the definition of a supervised strategy for ASIUM. Performances are then largely better, provided that the filter (i.e. the set of words initially defined by the analyst) is relevant. In function of the semantic class that the analyst wants to model, the filter has to change. There is here a subjectivity factor in our evaluation: elaborating a filter is, by definition, a subjective task that influences the quality of the results.

Compared to the experiments in an unsupervised mode, the elaboration of keyword filters makes things work better. However, the reduced size of corpus does not make it possible to obtain really accurate base classes. Filters are simple list of relevant words: this resource is not informative enough to obtain really accurate results. Different experiments

(Nazarenko *et al.* (2001); Morin et Jacquemin (1999)) have shown that using more structures data allows to guide and refine the overall results. This kind of techniques leads to a better quality.

3.3 Criteria for the elaboration of semantic classes

Class modeling necessitates an important part of manual work. Even with a file of words used as a filter, the number of base classes is important, varying from 41 to 576 in function of the total number of words contained in the filter.

Most of the time, the nearest base class defined contained less than 30% of the desired elements. On the other hand, it is very rare to see more than 30% of a base class validated by the analyst.

The amount of work to provide in order to clean and aggregate the different class is then relatively high. It is not infrequent to see a final semantic class that was initially split into 20 base classes. The number of aggregation to validate is proportional. This point can be problematic: the analyst must accept to spend some time to validate classes that are initially of very poor quality.

The classes obtained are progressively validated and refined. They are initially noisy but the validation phase allows then to access to proposal from the system that are based on relevant classes that are more homogeneous. The task is finished when the system has no new proposal to do to the analyst. In our experiments, the validation phase rarely concerned more than 3 or 4 level of the ontology. If the results of the learning phase are too poor, the expert may decide to relaunch a learning process with a lower threshold. The risk is then to obtain more noisy classes. The choice of a threshold is still an empirical task and Asium does not provide any means to fix it in as assisted manner.

3.4 Measuring the gain brought by a learning mechanism for an automatic resource acquisition process

Our aim in this experiment is to acquire semi-automatically thanks to Asium, semantic classes from a corpus. In this framework of this experiment on the FirstInvest corpus, we evaluate the results by comparing them with those previously manually obtained by an expert.

The results of ASIUM take into account a frequency dimension. That is the reason why most of the major elements (the set of words that are relevant for the task and that appear frequently in the corpus) have been found. There are a few missing elements but they have a limited influence on the overall results, like *échange*, *émission*, *transaction*. We must however notice that some elements like *vendre* have not been found, even if they are common in our corpus. Lastly, the distributional analysis makes it possible to find some elements that have not previously been retained by the expert working manually (*désengagement*, *regroupement*, *scission*). The distributional analysis gives a more accurate and more global view of the corpus, allowing to take better decisions. Relevant elements that did not appear in the corpus are, of course, not found in the classes proposed by ASIUM.

Once they have been refined and validated, the semi-automatically defined semantic classes obtain good overall results. The corpus is very homogeneous: even if its size is reduced (243 texts, 45.334 words). The training corpus allows to obtain very good performances over the test corpus. In this context, using ASIUM together with an important validation phase allows to accurately cover the training (and the test) corpus.

From this tool, one can obtain simultaneously some lists of relevant nouns together with associated verbs. There is a real gain when compared to the manual reading of the corpus. The evaluator estimates that about 10 hours have been spent to defined semantic classes with ASIUM, to be compared with the 50 hours spent to read the corpus and manually elaborate the resources. Guiding the learning process with a set of initial keywords (supervised learning) allows to directly obtain relevant classes. This last point is a great advantage compared to the previous experiment, based on an unsupervised learning phase.

4 Using general linguistic resource: the semantic network from Memodata

We contrast the experiment previously done by using ASIUM with the manipulation of a semantic network, directly providing semantic classes. We used the semantic net from Memodata: this company has developed since

ten years a very compete network, the INTEGRAL DICTIONARY (le DICTIONNAIRE INTÉGRAL in French), together with a set of functionality and a programming interface (API). See (Dutoit, 2000).

4.1 The semantic network: THE INTEGRAL DICTIONARY and associated tools

THE INTEGRAL DICTIONARY is the name of the semantic network developed by the French company Memodata. This network is inspired from the Text-Sense Theory from Igor Melč'uk. The theoretical basis of the dictionary is the set of different senses associated with a word (a string of characters). The network is made of 186.000 word senses. From this point of view, the coverage of the network is comparable with that of Wordnet (Fellbaum, 1998). However, THE INTEGRAL DICTIONARY is very different when one observes its overall structure. The links between words are not only *is-a* links, they are not based on psychological assumptions. Instead, the analysis is based on a decomposition of words in a set of basic features called *sèmes* (in the framework of a componential analysis) that are structured through typed links.

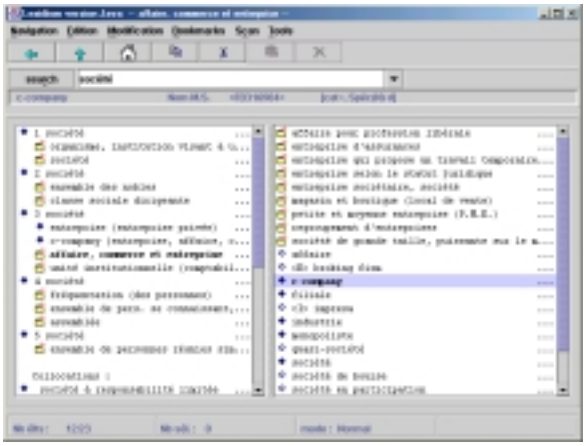

Figure 2. The INTEGRAL DICTIONARY

The network is structured thanks to an important set of links between elements of different syntactic category, like in EUROWORDNET (Vossen, 1998). This aspect is especially important for the task we are interested in, given that we must find that the French words *achat* (a noun) as well as acheter (a verb) reflect the notion of *acquisition*. The coverage of the INTEGRAL DICTIONARY is larger than the French EUROWORDNET. By the way, a part of the information contained inside the French EUROWORDNET has been automatically derived from MEMODATA's resources (the rest of the French EUROWORDNET network has been produced by the University of Avignon, that was responsible of the French part of the project).

4.2 Criteria for the elaboration of semantic classes

To elaborate semantic classes from a semantic network like the one from MEMODATA, the end-user must choose a keyword from the domain he is interested in and then query the linguistic base. A semantic class then progressively emerges when following the different links between elements in the semantic net.

In our experiment, we took into account all the words that were linked to the original keyword with one of the following typed links: *generic*, *specific* or *synonym*. The lexical class we obtain is then refined, so that we finally obtain an homogeneous and relevant semantic class (the analyst must essentially delete terms that are not relevant for the task or the domain. For example, in French, if the entry point (the initial keyword) is *achat* (*purchase*), the system will propose the following list of terms: *abonnement, accaparement, acheter, acquérir, acquisition, appropriation, commande, prise de contrôle, prise de participation, préemption, téléachat…* In this context of company purchasing other companies, the words *abonnement* and *téléachat* are irrelevant. They have to be manually discarded.

When we compare this approach with what we have previously done with ASIUM, we see that the end-user has no information about the representativeness of a word: the number of occurrences in the training corpus cannot be accessed. For the analyst, going back to the corpus to check the context is a too heavy and boring task. We will see in the next section that this can cause problem when dealing with the representativeness of the results.

The process stops when the analyst has checked the links between words and when no new term is found out. We observed that people rarely validates terms that are too much distant from the initial keyword. In fact, people rarely follow more than 2 to 3 nodes from the initial keyword.

In our experiments we noticed that the strategy we have adopted is efficient and allows to rapidly propose an accurate solution. The links that are followed by the analyst rapidly provide a set of already registered terms or some set of irrelevant terms. The acquisition stops when this kind of situations happens[2].

4.3 Measuring the gain brought by a general resources for the acquisition process

The evaluation of the INTEGRAL DICTIONARY for the task is very close from the one we have followed for ASIUM. We will compare the result of the acquisition process using the tool from MEMODATA with the semantic classes that have been manually defined.

However, we see a certain number of differences. The tool of Memodata does not necessitate any modeling task nor any training corpus. Even so, a certain number of subjectivity factors appeared during the evaluation. The analyst has to produce a request in order to obtain a set of related words. The starting point (the initial keyword chosen by the analyst) has a major effect on the quality of the overall results: one word can be more or less accurately linked with other words. The semantic network can be more detailed in one part than in some other parts, etc.

However, we noticed that, from a request made of 3 to 5 keywords, it is always possible to integrally reconstruct any semantic class whatever has been manually defined. For the evaluation, the difficulty consists, of course, in choosing the good initial set of keywords (like *achat* in the example of the previous paragraph), which are crucial for the quality of the results. This subjective factor in the evaluation is the same one as for ASIUM, when we had to define a set of keywords to define a filter. However, the small number of words that must be activated in the INTEGRAL DICTIONAY in order to be able to reconstruct the original classes shows the homogeneity of the semantic network from MEMODATA.

To avoid or, at least, limit this subjectivity factor, we chose to evaluate the system from a simple request consisting of a single keyword, instead of a set of keywords. We chose the keyword *achat* (*purchase*). The semantic class we obtained by activating the direct links of *achat* and then by validating these results, showed that more than 75% of the original class is found in this way (but we don't take into account the total number of occurrences in the corpus; we only calculate here the number of items appearing in the manually defined semantic class). 60% of the elements proposed by the Integral dictionary have been positively validated.

A large part of the class we obtained was not present in the corpus and, thus, was lacking in the class obtained manually or with ASIUM). We evaluated that about 40% of the elements proposed by the INTEGRAL DICTIONARY were lacking in the class defined manually although they were relevant elements. In other words, the semantic network from MEMODATA provides lots of relevant elements that were not in the training corpus and that the analyst could have forgotten.

However, things slightly change when we take into account, for each word, the total number of its occurrences in the corpus. We then observe that some key elements of the target semantic class are not found. The scores show that only 45% of the original semantic classes are found with the INTEGRAL DICTIONARY, when we take into account their total number of apparition in the corpus. The new elements proposed by the INTEGRAL DICTIONARY only improve the result by 7%. In other words, a lot of potentially relevant words are proposed that could be interesting, but these words are rarely present in the test corpus.

We should not conclude too quickly that general resources are inadequate for the task. A word like *raid* only appears once in the test corpus. But a research on a larger corpus from FIRSTINVEST showed that this term is relevant and productive in other news stories. Thus, we have to say that a "chance factor" is present: the nature of the corpus, its genre, the kind of language used are important factors that have an influence on the evaluation. The structure of the network and its homogeneity are also important factors. For example, if we access to the semantic network with the words *"achat"*, *"vente"* and *"fusion"*, we obtain more than 90% of the element of the manually defined class

[2] This is not obvious! The links that refer to Synonymy in a semantic network are rarely reciprocal (for example, a link can be established from A to B, but not from B to A), contrary to an open idea.

(more than 95% if we take into account the number of words really present in the corpus).

The conclusion for this part of the study is that the semantic classes proposed by the INTEGRAL DICTIONARY are of better quality than those of ASIUM. They cover a larger spectrum of elements since they are not limited to the training corpus. Unfortunately, some key elements are forgotten (30% of undiscovered elements correspond in fact to 55% of the occurrences of the corpus). The missing elements are very frequent ones and they contribute to a lower recall.

5 Conclusion: evaluating resource acquisition tools

Information extraction (IE) systems offer clear evaluation methods: a reference can generally be manually established and the results of automatic IE systems are compared with the reference. Things are very different when we try to evaluate resource acquisition tools.

Manually defined resources define a reference relative to a corpus and this reference is known to be largely imperfect. The evaluation is then not based on an objective and indisputable basis. These difficulties are increased again when we think to the fact that resource acquisition cannot constitutes an isolated and abstract task. Any linguistic modeling task requires a large knowledge of the domain and of the corpus. This knowledge allows the analyst to guide the machine learning process in defining good filters and good keywords. But, every part of human knowledge includes a part of unvoiced feeling and of implicitness that can hardly be described in an abstract model!

This part of subjectivity is inherent to the task and should not be hidden during the evaluation. We obtain results that are less readable than when we strictly follow a black box evaluation protocol. This lack of clarity can be compensated by the intuition of the expert who knows how to use his knowledge to appropriately use different acquisition tools. He is then able to develop original analysis strategies to explore the corpus and extract the good information.

It is then necessary to provide guidelines associated with acquisition tools. These guidelines should propose a method and a way to acquire knowledge from a given corpus. They should take into account the corpus size, the regularity of its vocabulary and of its syntax, and of its representativeness of a domain.

6 Bibliography

DUTOIT D. – *Quelques opérations texte → sens et sens → texte utilisant une sémantique linguistique universaliste apriorique*. Thèse de Doctorat en Informatique, Université de Caen, 2000.

FELLBAUM S. (éd.)– *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, 1998.

HIRSCHMAN L., GRISHMAN R. et SAGER N. – « Grammatically-based automatic word class formation ». *Information Processing and Management*, vol. 11, 1975, pp. 39–57.

GRISHMAN R. et STERLING J. – « Generalizing automatically generated selectional patterns ». In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING'94), Tokyo, 1994, pp. 742–747.

HARRIS Z. – Structural Linguistics. University of Chicago Press, Chicago, 1951.

NAZARENKO A., ZWEIGENBAUM P., BOUAUD J. et HABERT B. – « Corpus-based identification and refinement of semantic classes ». *Journal of the American Medical Informatics Association*, vol. 4, 1997, pp. 585–589.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. et BOUAUD J. – « Corpus-based extension of a terminological semantic lexicon ». In BOURIGAULT D., JACQUEMIN C. et L'HOMME M.-C. (éds.) – *Recent advances in Computational Terminology*. John Benjamins, Amsterdam, 2001, pp. 327–352.

FAURE D. – Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de texte : le système ASIUM. Thèse de Doctorat en Informatique, Université Paris 11–Orsay, 2000.

FAURE D., NÉDELLEC C. – « ASIUM: Learning subcategorization frames and restrictions of selection ». In *Proceedings of the ECML'98 Workshop on Text Mining*, Chemnitz, 1998.

FAURE D. et POIBEAU T. – « Extraction d'information utilisant INTEX et des connaissances sémantiques apprises par ASIUM, premières expérimentations ». In *Actes du 12ème Congrès de Reconnaissance des Formes et Intelligence Artificielle* (RFIA'2000), Paris, 2000, pp. 91–100.

MORIN E.and JACQUEMIN C. – « Projecting corpus-based semantic links on a thesaurus ». In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics

(ACL'99), Université du Maryland, 1999, pp. 389–396.

VICTORRI B. – « La construction dynamique du sens ». In Actes du 11ème Congrès de Reconnaissance des Formes et Intelligence Artificielle (RFIA'1998), Clermont Ferrand, 1998, pp. 15–29.

VOSSEN, P. (éd.) – EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht, 1998.