Automated Information Extraction out of Classified Advertisements*

Ramón Aragüés Peleato, Jean-Cédric Chappelier and Martin Rajman

Computer Science Dep. – Swiss Federal Institute of Technology (EPFL) – Lausanne

Abstract. This paper presents an information extraction system that processes the textual content of classified newspaper advertisements in French. The system uses both lexical (words, regular expressions) and contextual information to structure the content of the ads on the basis of predefined thematic forms. The paper first describes the enhanced tagging mechanism used for extraction. A quantitative evaluation of the system is then provided: scores of 99.0% precision/99.8% recall for domain identification and 73% accuracy for information extraction were achieved, on the basis of a comparison with human annotators.

1 Introduction

The work reported in this paper has been carried out in the context of the development of a system able to automatically extract and structure information from the textual content of newspaper advertisements. The system consists of three modules, as summarized in figure 1:

1. The task of the first module is to classify advertisements into *a priori* known classes (*real estate*, *vehicles*, *employment* or *other*). This step is needed to identify which thematic form has to be associated with the advertisement, and then used to guide the information extraction process. Classification is performed using a mixture of a naive Bayes classifier and a form-based classifier developed in our laboratory [15]. An evaluation on a test collection of 2,856 manually classified ads produced the very satisfying scores of 99.8% recall and 99.0% precision.
2. The task of the second module, which represents the main focus of this paper, consists in tagging (i.e. labelling) the textual content of the advertisement, in order to identify the information units that have to be extracted to fill in the slots of the associated form. Tagging is achieved by using specialized lexica, regular expressions, word spotting techniques and relative position analysis as described in the following sections.
3. Finally, the structuring module is in charge of transforming the tagged text into structured data (i.e. a filled form). This involves extracting the tagged

* Appears in 5th International Conference on Applications of Natural Language to Information Systems (NLDB'2000), Versailles (France), june 2000.

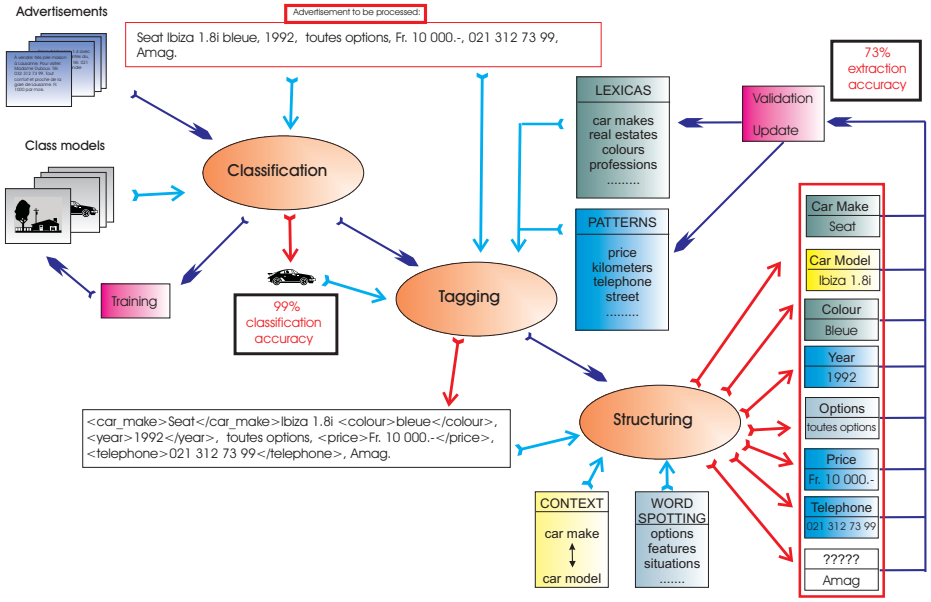

Fig. 1. *Global architecture of the system for automatic processing of newspaper advertisements.*

textual units, standardizing formulations[1], removing inappropriate punctuation, transforming abbreviations, etc. In the current system, this module remains quite simple as the tags used in step 2 closely correspond to the slots present in the associated forms.

The tagging phase can be further decomposed in the following steps:

- Labelling known entities (words, compounds, expressions) using specialized lexica and regular expressions (section 2).
- Identifying the nature of the information that is expressed by the textual units that have not been tagged in the first step (section 3). This is achieved through:
  1. segmentation based on punctuation and prepositions;
  2. word spotting in each segment (subsection 3.1);
  3. contextual tagging using the relative position of the units with relation to already tagged segments (subsection 3.2).

Notice that the design methodology used for our system is different from typical Information Extraction approaches [9] in the sense that, instead of trying to find some specific information in a whole document, it rather tries to identify the nature of the information expressed by each single piece of the text. In addition, the general strategy used by traditional systems [1, 8] consists in searching

[1] for example, using the same format for all price indications.

trigger words and then analyzing their context, while our system first segments using known entities and then analyzes the unknown segments with positional techniques and trigger words. Another specificity of our system is the average length of the processed documents: advertisements are generally short and very concise.

2 Tagging with Lexica and Regular Expressions

As already mentioned, the first step necessary to extract information from advertisements and fill in the automatically associated forms consists in tagging the advertisements for known entities using both specialized lexica and regular expressions [7].

2.1 Lexicon

Lexicon-based tagging simply consists in searching the text for entries contained in an *a priori* build lexicon. This lexicon may contain general words (e.g. *camion* [truck]), specific words (e.g. *airbag*), compounds (e.g. *pneus d'hiver* [winter tires]) and expressions (e.g. *libre de suite* [vacant immediately]) associated with identification labels (e.g. label such as *ville* [city] for the word *Paris*). Elements in the advertisement are tagged with the corresponding label only if they are non-ambiguous in the lexicon (i.e. associated with only one single label).

The tagging lexicon used in our system was created on the basis of a preliminary lexical study of a corpus of 10,700 advertisements, spread over 8 years[2]. A frequency analysis of the vocabulary was performed to serve as a guideline for the creation of the lexicon. For this analysis, a general purpose French lexicon containing more than 550,000 word forms (84,000 lemmas) was used and the following two questions were addressed:

1. what is the overall orthographic quality of the advertisements? The answer to this question determines whether an efficient spelling checker needs to be integrated in the system.
2. what is the proportion of specific vocabulary (i.e. vocabulary that is frequently used in advertisements but unknown to the general purpose lexicon)?

To answer these questions, the following table was built for the identified out-of-vocabulary forms (7270, 38.8% of the vocabulary):

		rare forms	frequent forms
Corrected	short	501 (2.7% of voc.)	514 (2.7% of voc.)
	long	1038 (5.5% of voc.)	
Not corrected		4170 (22.3% of voc.)	1047 (5.6% of voc.)

[2] The total vocabulary contained in that corpus was of 18,720 words. Words had an average frequency of 22 and advertisements had an average length of 37 words.

Rare forms are the forms[3] that appeared less than 3 times in the corpus. *Corrected forms* refer to forms that accept a one spelling error correction[4] in the general purpose lexicon (short/long refers to the number of characters in the form, short standing for less than or equal to 4 characters).

To interpret the above table, the following hypotheses were used

- frequent out-of-vocabulary forms that are not corrected correspond to instances of the specific vocabulary for the advertisements;
- frequent forms that can be corrected should be carefully analyzed as they might either correspond to systematic errors (frequent) or specific vocabulary that incidentally also corresponds to a correction that belongs to the general purpose vocabulary;
- rare and corrected forms may possibly be spelling errors. This has to be moderated by the length of the form as short forms more easily produce one spelling error corrections in a general purpose lexicon. We therefore decided to only trust corrections for forms with length greater than 4. Short rare forms are ignored, even if they have a correction in the general purpose lexicon.[5]
- rare and uncorrected forms are ignored as they concern infrequent phenomena for which not enough information is available.

With such interpretation rule the table can then be summarized as:

	rare forms	frequent forms
Corrected $\frac{\text{short}}{\text{long}}$	ignored spelling errors	manual processing
Not corrected	ignored	specific voc.

The above results therefore indicate that the corpus is of good orthographic quality (38.8% of out-of-vocabulary forms among which only 5.5% can reasonably considered as errors)[6] and contains a quite high ratio of specific forms (5.6% of identified specific vocabulary and 25%[7] of ignored forms mainly due to rare (personal) abbreviations).

The lexical study was also a good starting point for the creation of the tagging lexicon: first of all, many of the frequent unknown words were directly introduced into the lexicon, thus improving its coverage; but, most of all, all the new words identified were good indicators of what was the kind of vocabulary that can be found in newspapers advertisements. Therefore, when a word corresponding to a label was added to the lexicon (e.g. *Paris* being a city), several

[3] i.e. tokens resulting from a French tokenizer, most often words.

[4] a "one spelling error correction" is any form in the lexicon that is at an edit distance equal to 1 from the considered out-of-vocabulary form.

[5] This choice is sensible if the orthographic quality of the corpus is good, as it was the case for us.

[6] This was not surprising as we were dealing with proof-read newspapers advertisements. The results would certainly have been different if dealing with Internet advertisements.

[7] 22.3% + 2.7%

other words corresponding to the same label (e.g. all cities in the considered country) were also added. These other words have been extracted from several different sources, mainly Internet public lists. However, a large amount of time needed to be devoted to the validation/correction of these other sources of information. Approximately 45 person·days were spent on the lexical analysis and lexicon construction.

The following table shows examples of labels contained in the tagging lexicon used by our system:

Label	Number of words	Examples
Cantons	97	*GE, Genf, Genève*
Colours	36	*blanche, foncé, métallisée*
Car makes	134	*Renault, Seat, VW*
Car garages	114	*Amag, Croset, ROC*
Real Estates	137	*Chapuis, Gérim, Rêve-Immob*
Professions	539	*pompier, ingénieur, serveuse*
Languages	47	*Espagnol, Anglais, Roumain*
Months	24	*Janvier, Janv., Juin*
Motor bikes	48	*Honda, Yamaha, CBR*
Streets Index	83	*Rue, Av., Ruelle*
Cities	4230	*Lausanne, Zürich, Chur*
Kinds of vehicles	12	*Scooter, Bus, camion*
Kinds of buildings	57	*Halle, Appartement, villa*
Salary expressions	10	*salaire à discuter*

2.2 Regular Expressions

The second method used for directly tagging textual units was to apply descriptive patterns written with regular expressions, as for example dates, phone numbers, prices, surfaces. In order to create the regular expressions a first basic set was build for several *a priori* chosen slots of the forms to be filled. The resulting tagger was then run over a training corpus consisting of textual units corresponding to the chosen slots. New patterns were then gradually created and old ones improved by iterative testing on the reference corpus as long as there were slots with error frequency greater than 1.

The following table describes several different patterns created with this procedure:

Pattern	Examples
Name	*M. Duboux*
Surface	*200 m2 environ*
Kilometers	*100 000 km*
Number rooms	*3 1/2 pièces*
Age	*Agée de 35 ans*
Work Time	*50% ou 75%*
Action	*Cherche à louer*
Date	*Janvier 2000* *1.12.1999*

Pattern	Examples
Free	*libre de suite* *livrable: fevrier 2000*
Email	*pepito.grillo@cdi.com*
Price	*Fr. 3'000.- à discuter* *loyer à négocier*
Charges	*Charges comprises* *+ 50.- charges*
To visit	*Pour visiter: 021 693 66 97*
To treat	*Rens: Régie Houx, 021 693 66 97*
Telephone	*Tél. (021) 312 73 99, le soir.*

2.3 Tagging Known Entities

Using the above described tagging lexicon and regular expressions, the system then scans the whole advertisement and tags all the identified unambiguous textual units with the corresponding label. The output of this process therefore consists in a partially tagged text with remaining untagged parts corresponding to either unknown ambiguous items. Figure 2 gives an example of the result of this first step on a vehicle advertisement.

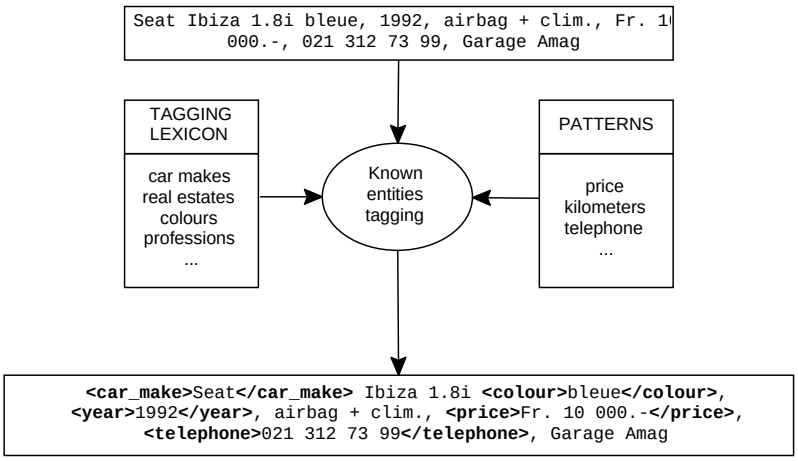

Fig. 2. *Tagging known entities with lexica and patterns. Information is identified by the SGML surrounding tags.*

3 Identifying Information in Unrecognized Parts

Once the advertisement has been tagged for known entities, it still contains several portions that have not been identified (e.g. *Ibiza*, *Garage Amag* in the

example of figure 2). To further tag these pieces of text the following three steps were applied:

1. the untagged text is segmented using punctuation and (for employment advertisements only) prepositions, so as to separate different information pieces that may be contained in the same text area[8]. A special treatment using a list of known abbreviations avoids segmenting punctuation used for abbreviations.
2. a word spotting score is computed for each segment on the basis of several trigger lexica (section 3.1).
3. If the word spotting score is not high enough to allow a reliable decision, the segment is tagged according to contextual rules taking into account the nature (i.e. the tags) of its neighbour segments (section 3.2).

3.1 Word Spotting

To compute the word spotting score for a segment, the system uses several *trigger lexica*. A trigger lexicon consists in a list of keywords that are typical for a certain type of information (e.g *climatisation* for vehicle options) and that help to identify the proper label for all the text in that segment. The word spotting score is therefore a measure of the likelihood for a segment to be relevant for a certain type of information [11].

The words selected for the trigger lexica used by our system were extracted from the lexical study described in section 2. They have been extended by running the system over a training corpus containing additional advertisements. Notice however, that trigger lexica differ from the tagging lexicon in the sense that they do not contain words that represent alone an identified information entity. These words rather give an idea on the kind of information that is contained in the text area they appear in.

The following table describes several different trigger lexica used in our current system:

[8] These segments may be recombined afterwards (at the end of the process) if they happen to have the same final tag.

Trigger lexicon	Number of words	Examples
Options (vehicles)	83	*autolook, clim, alarme*
Models (vehicles)	500	*Mégane, Punto, Ibiza*
Dealers (vehicles)	3	*garage, SA, AG*
Price (vehicles, real estate)	13	*prix, CHF, gratuit*
Construction (real estate)	11	*récent, rénover, refait*
Features (real estate)	42	*cheminée, balcon, parking*
Quality (real estate)	11	*spacieux, splendide, charmant*
Situation (real estate)	44	*calme, gare, centre*
To treat (real estate)	12	*renseignements, SARL, vente*
To visit (real estate)	3	*visite, contact, adresser*
Activity (employment)	70	*garder, ménage, nettoyage*
Age (employment)	3	*ans, âge, adulte*
Contact (employment)	6	*offre, contact, soumettre*
Qualifications (employment)	40	*diplômé, connaissances, programation*
Requirements (employment)	15	*curriculum, permis, véhicule*
Salary (employment)	3	*argent, gagner, salaire*
Company (employment)	15	*institution, SA, S.A.R.L.*
Work place (employment)	50	*pizzeria, hôtel, commerce*

As each trigger lexicon is associated with a unique specific tag, the word spotting score for each tag is computed as the number of words of the corresponding lexicon that appear in the segment. Finally, if there is a word spotting score that exceeds the others of a given threshold, the segment is tagged with the label associated with the corresponding trigger lexicon.

3.2 Contextual Tagging

In case where word spotting techniques do not permit to identify the information contained in a segment[9], a tag is allocated to the segment on the basis of the tag immediately preceding (i.e. its left boundary). We call this technique *contextual tagging* as the allocated tag to a segment depends on its (left) context. For example, the contextual tag following the make of a vehicle is "*model*" as, in vehicle advertisements, the car model very often follows the car make. This relation between tag and context is based on an *a prior* analysis carried out on a large amount of advertisements. When no contextual rule can be applied, the segment is tagged as "*undefined*".

The final result of the tagging module is therefore a fully tagged text that can then be directly used to fill the associated form. In the current system the structure obtained by filling the slots is further filtered: unwanted punctuation is removed and slots without relevant information (i.e. less than one normal[10] character) are removed.

[9] i.e. all word spotting scores are under the threshold or there is more than one best score (ties).

[10] neither blank nor punctuation.

3.3 Examples

The following table contain several examples of tags obtained by word spotting and contextual tagging:

Class	vehicles		real estate		employment
Tag	options	models	features	situation	qualifications
Vocabulary size	291	266	2586	1169	522
Examples	pneus neufs 3 portes toit ouvrant 5 portes ABS jantes alu radio CD toutes options climatisation	A 160 Avantgarde Astra Break 16V 4x4 LX (241 HSE) Grand Cherokee	meublé jardin 2 salles d'eau place de parc garage cave balcon cuisine agencée	centre 1er étage situation calme 2e étage 3e étage vue imprenable vue sur le lac vue calme	sérieux dynamique jeune fille jeune homme d'expérience avec expérience jeune

Underlined words are words that appear in trigger lexica (and were used for word spotting).

An example of the the final output of the tagging module for the same vehicles advertisement as in figure 2 is given in figure 3, as well as its corresponding final filled frame.

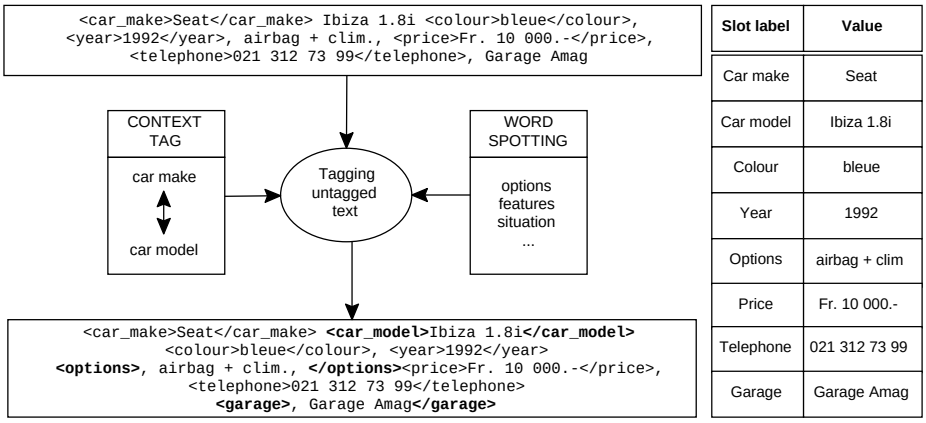

Fig. 3. *Tagging entities with word spotting and context. The frame filled after structuring is also shown on the right.*

4 Evaluation of the system

An evaluation of the system described in this paper was done on the basis of a comparison with forms filled by human annotators.

The test corpus consisted of 77 advertisements (41 real estate, 30 employment and 6 to vehicles; a proportion representative of both the whole corpus and week

ads production). Each of these advertisements was submitted [11] to several human annotators who had to properly fill the corresponding form.

Before the evaluation of the system itself, the manually annotated forms were checked for coherence using a kappa measure [18, 3, 2]. On the basis of the confusion matrices produced for the computation of the kappa scores, several unreliable slots were thus identified and were then either removed or clustered in order to improve the agreement among the human annotators. The final average kappa value obtained was 0.9, thus indicating a satisfying agreement for the reference[12].

The test set was then submitted to the system, and the results were compared to the human references. The rules used to assign a comparison score to a slot were the following:

- If there is no agreement among the manual annotation, the slot is ignored (neutralization (`NTR`) case);
- If no value, neither manually nor automatically, was assigned to the slot, it is ignored (non evaluation (`NEV`) case);
- If both the system and the human annotators agree, the answer is considered as correct (`OK` case);
- In any other case the answer is considered as an error (`ERR` case).

Notice that the evaluation of the slots was carried out manually and the `OK`/`ERR` assessments for the values assigned to the slot by the system and the annotators were therefore judged by humans. On a larger scale, the manual assessment would need to be replaced by automated rules relying for instance on the number of common words in the values assigned.

The global accuracy of the system is then measured on all cases where a coherent answer was provided by human annotators (i.e. on all `OK` and `ERR` cases) by the ratio `OK`/(`OK` + `ERR`).

The test corpus of 77 advertisements contained 1415 different slots[13] among which 556 where actually filled by human annotators and 519 exhibited sufficient agreement for the manual annotation. Among these 519 slots, the system provide a correct value in 381 cases, leading to a global accuracy of 73% correct extractions (70.5% for employment, 73% for real estate, 88% for vehicles).

Further results, detailed by domains and by slots, are given in table 4.

5 Conclusions and Future Work

The goal of the work presented in this paper was to create a system able to automatically classify and structure newspaper advertisements. As the classified advertisements domain is quite different from other studied information extraction problems [12–14] specific techniques were implemented, which as shown in

[11] together with its corresponding form

[12] kappa measure varies between -1 (complete disagreement) and 1 (complete agreement). Values higher than 0.8 are usually considered as indicating good agreement.

[13] 30 x 18 for Employment, 41 x 19 for real estate and 6 x 16 for vehicles.

Vehicle	OK	ERR	NTR	NEV	Score (%)
Action	2	0	0	4	100
Kind of vehicle	0	1	2	3	0
Vehicle make	4	0	1	1	100
Vehicle model	4	1	0	1	80
Motor bike	0	0	0	6	–
Colour	3	0	0	3	100
Year	5	0	0	1	100
Expertized	2	0	0	4	100
Kilometers	5	1	0	0	83
Options	4	0	1	1	100
Price	3	1	0	2	75
Dealer	0	1	0	5	0
Contact	0	0	0	6	–
Telephone	6	0	0	0	100
Fax	0	0	0	6	–
E-mail	0	0	0	6	–
Total	38	5	4	49	**88.4**

Real Estate	OK	ERR	NTR	NEV	Score
Action	31	1	2	7	97
Kind of building	28	6	4	3	82
Number of rooms	16	3	1	21	84
Surface	10	2	2	27	83
Story	8	21	6	6	28
Construction	1	3	2	35	25
Features	17	7	2	15	71
Region	24	14	2	1	63
Quarter	0	4	1	36	0
Street	2	5	1	33	29
Price	22	0	0	19	100
Charges	5	4	0	32	56
Entry date	15	4	0	22	79
Real estate	0	1	0	40	0
To visit	0	1	0	40	0
To treat	2	1	1	37	67
Telephone	33	2	1	5	94
Fax	0	0	0	41	–
E-mail	0	0	0	41	–
Total	214	79	25	461	**73.0**

Employment	OK	ERR	NTR	NEV	Score
Action	30	0	0	0	100
Job	17	9	2	2	65
Qualifications	9	13	0	8	41
Age	0	4	1	25	0
Languages	1	1	0	28	50
Entry date	16	2	0	12	89
Work time	3	1	2	24	75
Kind of Company	0	8	0	22	0
Work place	9	8	1	12	53
City	12	1	1	16	92
Quarter	0	0	0	30	–
Street	0	1	0	29	0
Region	3	2	0	25	60
Salary	0	1	1	28	0
Contact	10	3	0	17	77
Telephone	18	0	0	12	100
Fax	1	0	0	29	100
E-mail	0	0	0	30	–
Total	129	54	8	349	**70.5**

Table 1. Precise results of the evaluation of the system for each slot of the 3 forms corresponding respectively to real estate, employment and vehicle advertisements.

section 4 achieve very promising results (73% correct extraction) when compared with human annotators.

However, there is still room for improvements. In particular the presented tagging methodology has one important limitation: when the text that remains untagged after the first segmentation[14] contains information associated with different labels[15], the word spotting technique does not correctly tag the text. Indeed when the untagged text contains keywords for two different trigger lexica[16], a decision about the contents of that information unit is not possible (same

[14] where the system uses tagging lexicon and regular expressions

[15] e.g. "*situation calme et place de parc*" (calm and garage) contains information about the situation and the features of the building"

[16] e.g. keywords for "situation" and "parking place";

score for two trigger lexica means no decision about the kind of content) and the text is then tagged as undefined.

One way of solving this problem is to apply a progressive tagging, in which segmentation is not done on the sole basis of the tagging lexicon and patterns, but delayed until the nature of the information inside the segment is unambiguously identified. The idea is to progressively calculate the word spotting scores for a growing initial sequence of words in the untagged segment and to build a new segment (with an associated tag) only when the difference between the scores assigned by the different trigger lexica decreases. Experiments over the whole corpus of advertisements are being carried out, and future work will evaluate the potential improvements brought by this technique.

Another future research will focus on lowering the dependency on hand-written lexica and patterns. As shown in [10, 16, 17] different techniques allow a system to automatically extract patterns and dictionaries form labelled and unlabelled texts, allowing a faster adaptation of a system when moved to a new domain. Extending the approach to ontologies could also be considered [5, 4, 6].

References

1. D. Appelt et al. SRI international FASTUS system: MUC-6 results and analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6*. Morgan Kaufmann Publishers, 1995.
2. J. Carletta. Assessing agreement on classification traks: the kappa statistics. *Computational linguistics*, 2(22):249–254, 1996.
3. J. Cohen. A coefficient of agreement for nominal scales. *Educational and Phychological measurement*, 20:37–46, 1960.
4. D. W. Embley, D. M. Campbell, Y. S. Jiang, S. W. Liddle, Y.-K. Ng, D. Quass, and R. D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data and Knowledge Engineering*, 31(3):227–251, 1999.
5. D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured documents. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim, editors, *Proc. of of the 1998 ACM CIKM Int. Conf. on Information and Knowledge Managemen (CIKM'98)*, pages 52–59, Bethesda (Maryland), November 1998. ACM Press.
6. D. W. Embley, N. Fuhr, C.-P. Klas, and T. Rölleke. Ontology suitability for uncertain extraction of information from multi-records web documents. *Datenbank Rundbrief*, 24:48–53, 1999.
7. D. Fisher et al. Description of the UMASS system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Mateo, CA, 1995. Morgan Kaufmann Publishers.
8. R. Grishman. The NYU system for MUC-6 or where is the syntax? In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Mateo, CA, 1995. Morgan Kaufmann Publishers.
9. R. Grishman. Information extraction: Techniques and challenges. In ed. M. T. Pazienza, editor, *International Summer School SCIE-97*, Springer-Verlag, July 1997.

10. J.-T. Kim and D. I. Moldovan. Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE Transactions on Knowledge and Data Engineering*, 1995.
11. A. McCallum and K. Nigam. Text classification by bootstrapping with keywords, EM and shrinkage. In *ACL '99 Workshop for Unsupervised Learning in Natural Language Processing*, 1999.
12. *Proceedings of the Fifth Message Understanding Conference (MUC-5)*. Morgan Kaufmann, August 1993.
13. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, San Mateo CA, November 1995.
14. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/, 1997.
15. R. Aragüés Peleato, J.-C. Chappelier, and M. Rajman. Lexical study of advertisements and automatic identification of lexical items through contextual patterns. In *Proc. of 5th International Conference on the Statistical Analysis of Textual Data (JADT'2000)*, volume 1, pages 309–316, Lausanne (Switzerland), March 2000.
16. E. Riloff. Automatically generating extraction patterns form untagged text. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 50–55, San Jose, CA, 1992. AAAI/MIT Press.
17. E. Riloff. An empirical study of automated dictionary construction for information extraction in three domains. *Artificial Intelligence*, 85:101–134, 1996.
18. S. Siegel and N.J. Jr. Castellan. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, second edition edition, 1988.