Morphological resources for precise information retrieval

Anne-Laure Ligozat, Brigitte Grau, Delphine Tribout

Morphological resources for precise information retrieval

Anne-Laure Ligozat[1,2], Brigitte Grau[1,2], and Delphine Tribout[3]

[1] LIMSI-CNRS,F-91403 Orsay Cedex, France
[2] ENSIIE, 1 square de la résistance, 91000 Evry
[3] LLF, 5 rue Thomas Mann, F-75205 Paris Cedex 13

Abstract. Question answering (QA) systems aim at providing a precise answer to a given user question. Their major difficulty lies in the lexical gap problem between question and answering passages. We present here the different types of morphological phenomena in question answering, the resources available for French, and in particular a resource that we built containing deverbal agent nouns. Then, we evaluate the results of a particular QA system, according to the morphological knowledge used.

Key words: morphological resources, question answering

1 Introduction

Question answering (QA) systems aim at providing a precise answer to a given user question. Their major difficulty lies in the lexical gap problem: the answering document may not contain the exact same words as the question. QA and IR systems must find a way of retrieving relevant documents without relying only on mere identity between words. Linguistic knowledge must thus be used, and integrating morphological knowledge has often been preferred over semantics because the integration of morphological knowledge often is more reliable.

Most of the research carried out so far made use of simple heuristic-based stemming techniques which cut off word endings (such as [Lennon et al., 1988,Harman, 1991,Fuller and Zobel, 1998]). In most cases, the recall is slightly improved, but these techniques also produce some noise. Another way to use morphological knowledge is by extending the query, such as in [Moreau and Claveau, 2006], who significantly improve the results in most of the European languages for which they performed the experiment.

As we have shown, IR and QA applications mostly rely on partial or superficial morphological knowledge. However, some morphological resources are now able to provide detailed and precise knowledge about a large spectrum of morphological processes.

In this paper [4], we present the different types of morphological phenomena in QA, the resources available for French, and in particular a resource that we built

[4] This work was partly realized as part of the Quaero programme, funded by OSEO, French State agency for innovation.

containing deverbal agent nouns. Then, we evaluate the results of a particular QA system, according to the morphological knowledge used.

2 Morphological resources for QA

2.1 Morphological phenomena in QA

[Bernhard et al., 2011] studied the most frequent derivational relations between questions and documents in French. An answering document can contain various types of variants of question words: a possible answer to the question *When was dynamite invented?* is *Alfred Nobel is the inventor of dynamite and patented it in 1867*. In order to detect the correspondance between the question and this document, the relation between *invented* and *inventor* must be recognized.

[Bernhard et al., 2011] manually annotated several corpora of questions and answering documents with the types of morphological relations between them. They showed that the most frequent relations in open domain are flexional and derivational relations. Concerning derivations, the most common types are denominal adjectives (*région-régional* for *region-regional*), and nominalizations, in particular action nouns (*inaugurer-inauguration* for *inaugurate-inauguration*) and agent nouns (*réaliser-réalisateur* for *direct-director*).

In this work, we used these observations to integrate morphological knowledge into a QA system. To this end, we first considered the existing morphological resources for French. Two French morphological resources exist, for deverbal action nouns and relational adjectives, which we will now present.

2.2 Derivational resources for French

Verbaction [5] is a lexical resource containing action names derived from a verb [Hathout and Tanguy, 2002,Hathout et al., 2002]. It contains 9,393 noun-verb pairs, such as *renouveler-renouvellement* (*renew-renewal*).

Prolexbase [6] is a multilingual dictionary of proper nouns [Tran and Maurel, 2006,Bouchou and Maurel, 2008]. Although it does not contain explicit morphological knowledge, it gives information about relational nouns and adjectives derived from proper nouns. For example, the noun *Français* and the adjective *français* (*French*) are related to the entry *France*. Prolex contains 76,118 lemmas and 20,614 derivational relations.

Derivational resources thus exist for French, but concerning QA needs, a resource for deverbal agent nouns was lacking, so we built one semi-automatically, which we called VerbAgent, in reference to Verbaction.

[5] http://redac.univ-tlse2.fr/lexicons/verbaction.html
[6] http://www.cnrtl.fr/lexiques/prolex/

3 Construction and validation of a resource for deverbal agent nouns

As a first step, we automatically derived verbs from nouns, using formal properties of nouns. In French, some suffixes are frequently used to form deverbal agent nouns, such as *-eur*, as in *danseur* (*dancer*) derived from the verb *danser* (*dance*). Nine such suffixes were identified:

1. *-eur* (*danser* > *danseur*)
2. *-euse* (*chanter* > *chanteuse*)
3. *-rice* (*inspecter* > *inspectrice*)
4. *-eresse* (*défendre* > *défenderesse*)
5. *-aire* (*signer* > *signataire*)
6. *-ant* (*attaquer* > *attaquant*)
7. *-ante* (*diriger* > *dirigeante*)
8. *-ent* (*adhérer* > *adhérent*)
9. *-ente* (*présider* > *présidente*)

We then used a lexicon of inflected forms, Morphalou [7], to extract all nouns ending with one of the suffixes, and then checked if a corresponding verb existed in Morphalou. This verification was based on some twenty rules, such as: suppress the agent suffix *-eur* and replace it with the infitive suffix *-er*.

4,067 noun-verb pairs were generated. Yet, a formal resemblance between a noun and a verb does not guarantee that they are morphologically related: for example the pair *accentuer-accentueur* is extracted, although these two words are not morphologically related. Moreover, the noun and verb can belong to the same derivational family but without the noun being derived from the verb, such as in *rougir-rougeur*. A validation is thus necessary.

3.1 Manual validation

Manual validation consisted in checking that, for each considered pair, the noun was derived from the verb, and corresponded to an agent noun. We verified the semantic link or definitions in the TLFi [8] if necessary, for example when the noun was rare. 363 pairs were examined, among which 76% were correct i.e. contained a verb and the corresponding agent noun and 24% incorrect.

Errors mostly come from nouns with *-ant* or *-aire* suffixes, which can denote agents, but also commonly denote non agent nouns (such as *adoucissant*). Other errors come for example from nouns with an *-eur* suffix, which are frequently associated with an instrument (such as in *aspirateur*).

As this manual validation is very time-consuming and delicate, we defined several methods for an automatic validation.

[7] http://www.cnrtl.fr/lexiques/morphalou/
[8] http://atilf.atilf.fr/: on-line French dictionary

3.2 Automatic validation

We experimented several methods for an automatic validation of pairs.

Definition extraction from a dictionary First, we extracted the definition from the dictionary Littré [9]. Indeed, agent noun definitions usually begin with a phrase such as *Celui qui* (*The one who...*) followed by the corresponding verb. For example, the definition of the agent noun *chanteur* (*singer*) is *Celui, celle qui chante, qui fait métier de chanter* (*The one who sings, whose profession is to sing*).

Using the pattern *Celui, (celle)? qui* for detecting such definitions, we extracted 2,944 nouns. Yet, using only a pattern does not guarantee that the noun is derived from the verb. For example, it extracts the definition *Celui, celle qui joue du piano* (*A person who plays the piano*). Thus, we added a simple constraint on the verb form: the first two characters of the noun and the verb must be the same. This way, we extracted 1,121 nouns, which we hope to be more accurate, although less complete (for example the noun *agresseur* is not extracted because its definition is *Celui qui attaque le premier*, and does not contain the corresponding verb *agresser*).

We compared these lists of nouns to the manually annotated part of the resource. Using the definition pattern only, 92 annotated nouns are extracted, among which 87 were considered as actual deverbal agent nouns. Using the additional constraint on the verb, 60 nouns are extracted, which were all manually annotated as correct.

As 275 noun-verb pairs were manually annotated as correct, this automatic validation method does not present a very good recall (22% with the verb constraint); yet it is very precise. The low recall can be explained by the existence of different definition patterns (such as for *agresseur*), and by the absence of some rare words from the dictionary (such as *avaliseur*).

Cooccurrents In order to evaluate if two words are semantically related, it is also possible to rely on their contexts in corpus. Thus, we exploited a cooccurrence network [Ferret, 1998], extracted from a French corpus of articles [10]. Our hypothesis is that if a verb-noun pair shares cooccurrents, words in the pair are semantically related and more likely to be the result of a derivation.

We extracted, for each verb-noun pair, their closest cooccurrents [11], and considered that a pair was correct if it had at least one common cooccurrent. The main disadvantage of this method is the absence of many words, due to the

[9] XMLittré is an electronic version of the French dictionary Littré.

[10] This corpus was constructed based on 24 months of articles from Le Monde newspaper, using a 20 word window, and without taking order into account. Only cooccurrents with a frequency higher than 5 were kept. This network contains 31,000 words. Cohesion between words is based on mutual information estimation.

[11] All nouns were lemmatized according to the TreeTagger lemmatization usage, since this tagger was used to build the cooccurrence network.

limited size of the corpus: only 869 pairs are present in the network (both the noun and the verb exist) on the 4,067 pairs of the resource.

In order to test the relevance of this method, we compared the pairs presenting at least one common cooccurrent with the manually annotated part of VerbAgent. 85 annotated pairs are found in the cooccurrent network, among which 56 have a common cooccurrent, 45 of which being annotated as correct, and 11 as incorrect. Errors are usually semantically related pairs, but with a noun that does not correspond to an agent, such as *accablant-accabler* or *accélérateur-accélérer*. This method thus seems to give a clue on the relation between the noun and the verb, but a larger corpus would be needed to get more complete results.

N-grams In order to validate an equivalence of meaning of the two words, we exploited the distributional idea which states that related words share a same context. We considered a context made of one word, and defined rewriting rules for defining the usage of verb *vs* the usage of an agent noun. Our purpose is to recognize such rewritings: *chanteur d'opéra* (*opera singer*) vs *chanter un opéra* (*to sing an opera*) to validate the pair *chanteur-chanter* (*sing-singer*).

We used the Google Books Ngrams resource, which contains n-grams of words computed on digitized books. We collected n-grams which contain the nouns and verbs of VerbAgent, followed by either a determinant or a preposition plus a word. We considered that a pair is valid if both words share at least a same context, i.e. are used in relation with a same word. We extracted 1,795 n-grams corresponding to 231 pairs. Their evaluation on the reference set shows that within the 19 pairs found, 1 is not valid. This method seems to be precise; however it has a low recall.

Combination Table 1 presents the number of noun-verb pairs validated by each method, as well as by their combination. The second column indicates for example that 170 pairs were validated by all three methods, among which 15 were manually annotated as correct, and none was manually annotated as incorrect.

Table 1. # noun-verb pairs validated by each method

littré	*		*	*	*			
coocs.	*	*		*		*		
n-grams	*	*	*				*	
# found pairs	170	191	163	79	790	161	223	2290
correct	15	9	11	7	54	14	16	179
incorrect	0	10	0	0	5	1	12	61

The Littré seems to have the best recall, with a good precision. Yet, 2,290 pairs are found by none of the methods, mostly because the verb or noun frequencies are two low. A possible improvement could be to use larger or more

adapted corpora (for example results of Web queries using these words), or to use additional resources (such as other dictionaries).

4 Contribution of morphological knowledge to the answering process

4.1 General description of QAVAL

QAVAL [Grappy et al., 2011] is a QA system for French. Lucene is used to select shorts passages (instead of documents) which are then analyzed by a shallow terminological parser, Fastr [Jacquemin, 1999], for recognition of terms and their variants. Best passages are selected according to the presence of question terms. Candidate answers are then extracted from these passages and a machine learning validation system applies several criteria to rank the candidates.

4.2 Use of morphological knowledge in QAVAL

Morphological variants are handled at two stages in QAVAL: at passage retrieval and at passage selection as said before. At passage retrieval, the collection indexation and interrogation use stemming, which contributes to the presence of morphological variants of question terms.

4.3 Experiments

We conducted tests on two kinds of documents, Web documents and newspaper articles. We used 147 factual questions on a Web collection and 479 questions from CLEF and EQUER campaigns. To evaluate the impact of morphological resources, we calculated the MRR[12] on the first 10 passages selected for these questions, with and without morphological variants. We focused on the 10 best passages because after this rank, it is very difficult to extract a correct answer which would be proposed in the first ranks. Terms are searched in 150 passages retrieved by the search engine. After their annotation by terms which allows their weighting, we only keep the 50 best passages.

Results are presented in table 2. The first two columns indicate the collection which is searched, and the total number of questions studied. Column *#q SS* gives the total number of questions that can be answered without taking into account variations and *#q VAR* with variations. A question can be answered if it is associated with at least one passage containing the expected answer.

In order to compare the QA system performances under the same conditions, we determined the questions which can be answered (column *#q OK* table 2). Then, we kept among these questions those such as at least one passage contains variations of the question terms (column ExistVAR, subset of the column *#q OK*). Columns *MRR SS* and *MRR VAR* give the MRR computed on the passages associated to this last set of questions, annotated without and with variants respectively.

[12] Mean Reciprocal Rank: average of the reciprocal ranks of the first correct answers

Table 2. QAVAL results when selecting passages, with and without morphology

collection	#quest.	#q OK	#q SS	#q VAR	#q ExistVAR	MRR SS	MRR VAR
clef05	197	187	175	174	125	0.6298	0.6486
clef07	156	92	86	82	49	0.5269	0.5484
equer	126	117	105	105	96	0.6782	0.7039
quæro	147	125	106	113	76	0.3984	0.4347
total	626	521	472	474	346	0.5778	0.6027

The overall number of questions which can be answered does not vary in the two cases. However, differences come from the impact of variations in the ranking process. Adding morphological knowledge systematically improves the MRR, for each collection, and each question set. On all collections, the MRR without any morphological knowledge is 0.5785 and 0.6096 when taking into account morphological knowledge. This kind of improvement, even small, is important for a QA system. Extraction of answers is based on the capacity of a system to compare a question and an answering passage, especially when the wordings are different. A good matching will rely on resources having a good coverage.

QAVAL searches for all kinds of morphological variations. Thus, we evaluated which kinds of variants are found in all the passages retrieved by Lucene, and in the passages which contains a correct answer (see Figure 1). Following the study conducted in [Bernhard et al., 2011], we categorized variations in verb-action variants (*action* in the table), verb-agent (*agent*), location noun-adjective (*relat*) and others, as for example adjective-adverb etc. Percentages of each kind of variations are computed for questions coming from different QA campaigns and we can see that if *others* variations are numerous in all the passages, they are less important in the correct passages. On the other hand, *relat* and *action* variants are well represented in correct passages. We can also see that *agent* variants are less frequent; however these results have to be confirmed on a larger corpus. The cumulate results are given in the last two columns.

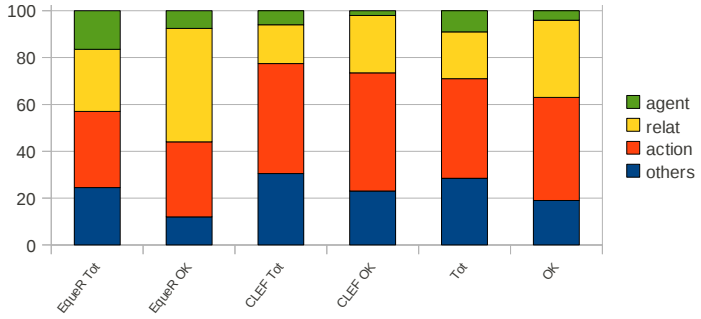

Fig. 1. Repartition of kinds of variants in passages (all passages: tot, correct ones: OK)

5 Conclusion

We presented in this paper a method for constructing a precise terminological resource containing morphological relations. This method leads to some errors and requires validation. Variant occurrences in corpora or resources were used to validate some relations automatically.

We also experimented using morphological resources in a question answering system, QAVAL, and found that such a knowledge leads to a better selection of passages, and that some kinds of morphological derivations are more frequent in passages which contain the correct answer than in other passages.

References

[Bernhard et al., 2011] Bernhard, D., Cartoni, B., and Tribout, D. (2011). A Task-based Evaluation of French Morphological Resources and Tools. *Linguistic Issues in Language Technology*, 5(2).

[Bouchou and Maurel, 2008] Bouchou, B. and Maurel, D. (2008). Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres. *Traitement Automatique des Langues*, 49(1):61–88.

[Ferret, 1998] Ferret, O. (1998). *ANTHAPSI : un systme d'analyse thmatique et d'apprentissage de connaissances pragmatiques fond sur l'amorage*. PhD thesis, Paris Sud.

[Fuller and Zobel, 1998] Fuller, M. and Zobel, J. (1998). Conflation-based comparison of stemming algorithms. In *Proceedings of the Third Australian Document Computing Symposium*.

[Grappy et al., 2011] Grappy, A., Grau, B., Falco, M.-H., Ligozat, A.-L., Robba, I., and Vilnat, A. (2011). Selecting Answers to Questions from Web Documents by a Robust Validation Process. In *Web Intelligence*.

[Harman, 1991] Harman, D. (1991). How effective is suffixing? *Journal of the American Society of Information Science*.

[Hathout et al., 2002] Hathout, N., Namer, F., and Dal, G. (2002). *Many Morphologies*, chapter An Experimental Constructional Database : The MorTAL Project, pages 178–209. Cascadilla Press.

[Hathout and Tanguy, 2002] Hathout, N. and Tanguy, L. (2002). Webaffix: Discovering Morphological Links on the WWW. In *Proceedings of the Third International Conference on L anguage Resources and Evaluation*, pages 1799–1804, Las Palmas de Gran Canaria, Espagne.

[Jacquemin, 1999] Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th annual meeting of ACL*.

[Lennon et al., 1988] Lennon, M., Pierce, D. S., Tarry, B. D., and Willett, P. (1988). An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*.

[Moreau and Claveau, 2006] Moreau, F. and Claveau, V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisime Confrence en Recherche dInformations et Applications CORIA 2006*.

[Tran and Maurel, 2006] Tran, M. and Maurel, D. (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, 47(1):115–139.