Classification de questions par traduction

Anne-Laure Ligozat

LIMSI-CNRS
rue John von Neumann

91403 Orsay Cedex
prenom.nom@limsi.fr

RÉSUMÉ. *Dans cet article, nous nous intéressons à la classification de questions pour un système de questions-réponses en français. Faisant le constat d'un manque de corpus annoté en français, nous nous interrogeons sur la possibilité d'exploiter les corpus anglais existants, en utilisant des traducteurs automatiques. Nous avons mené une série d'expériences en faisant varier le sens de traduction des corpus et les attributs fournis au classifieur. Les résultats montrent qu'il est possible de s'approcher des performances monolingues en traduisant le corpus d'apprentissage.*

ABSTRACT. *In this paper, we focus on question classification for a French question answering system. Since there are no annotated corpus for French, we investigated exploiting the existing corpora for English, by using machine translation tools. We conducted a series of experiments varying the translated corpora and the features given to the classifier. Results show that it is possible to obtain an accuracy close the monolingual one by translating the training corpus.*

MOTS-CLÉS : *Systèmes de questions-réponse, classification de questions, traduction automatique.*

KEYWORDS: *Question answering systems, question classification, machine translation.*

1. Introduction

Dans le domaine des questions-réponses, comme dans de nombreux domaines du traitement automatique des langues et de la recherche d'information, l'anglais est la langue la plus étudiée, et la plus dotée en termes de ressources, corpus et systèmes. De nombreuses méthodes sont fondées sur des apprentissages supervisés, qui bénéficient de la grande quantité de ressources et corpus disponibles pour l'anglais.

Lors du développement d'un système de réponse à des questions pour le français, nous nous sommes donc confrontés au problème du manque de ressources pour cette langue. Certaines ont été construites, par exemple pour l'apprentissage de la validation de réponses (Grappy *et al.,* 2011). Pour le module spécifique d'analyse de la question, nous avons développé deux systèmes pour le français : un premier fondé sur des règles (Ligozat, 2006) permettant de pallier l'absence de corpus de questions annotées, et un second, par apprentissage supervisé, pour lequel nous avons annoté un corpus d'environ un millier de questions. Cette annotation étant très coûteuse en temps, nous nous sommes également interrogés sur la possibilité d'utiliser les corpus anglais existants.

Le transfert de connaissances d'une langue à une autre se fait généralement à partir de corpus parallèles, mais dans notre cas, nous ne disposons pas de tels corpus. Nous avons donc étudié la possibilité d'utiliser des traductions automatiques pour créer un corpus de questions parallèles, comme cela a été fait par exemple en compréhension de l'oral (Jabaian *et al.,* 2011).

Cet article présente les expériences effectuées dans ce cadre. Dans la section 2, nous présenterons l'objectif de la classification de questions dans le cadre des systèmes de questions-réponses, et la façon dont elle a été mise en œuvre pour ce travail. Puis, nous présenterons les deux méthodes de passage d'une langue à l'autre que nous avons testées dans la section 3 : traduction du corpus d'entraînement ou traduction du corpus de test. Enfin, nous donnerons les résultats des expérimentations que nous avons menées section 4, et terminerons par un tour d'horizon des travaux similaires section 5.

2. Classification de questions pour un système de questions-réponses

2.1. *Système de questions-réponses QAVAL*

Un système de questions-réponses vise à répondre à une question posée en langage naturel par une réponse courte, au lieu d'une liste de documents. Ainsi, à la question «Quelle est la hauteur de la Tour Eiffel ?», un tel système va retourner la réponse précise «334m», accompagnée d'un passage justificatif.

Nous nous intéresserons ici plus particulièrement à la première étape d'un système de questions-réponses, c'est-à-dire l'analyse des questions.

2.2. *Analyse des questions*

Différents types d'analyse peuvent être effectués, mais la plupart des systèmes incluent une étape de classification des questions en fonction du type de réponse attendu. Ainsi, pour la question «Qui est le président des États-Unis ?», la réponse attendue est un nom de personne. Ce type sera ensuite utilisé lors de l'extraction de la réponse, afin d'extraire en priorité des noms de personnes comme réponses.

La détection du type de réponse attendu est généralement considérée comme un problème de classification multi-classes, chaque type de réponse possible représentant une classe. Afin d'apprendre à classer les questions, il convient alors de disposer de corpus de questions annotées, précisant le type de réponse attendu pour chaque classe.

2.3. *Paramètres de classification*

Dans notre système QAVAL, nous avons développé un système de classification de questions sur ce principe. Les attributs de base fournis au classifieur sont les n-grams de mots de la question.

Nous avons utilisé l'outil LibSVM (Chang *et al.,* 2011) comme classifieur, car il permet de faire de la classification multi-classes (en mode un-contre-un). Afin de pouvoir comparer les résultats obtenus notamment par (Zhang *et al.,* 2003), nous avons conservé les paramètres par défaut de cet outil.

2.4. *Classes de questions*

Concernant les types de réponses attendus, nous avons repris la taxonomie de questions proposée par (Li *et al.,* 2002) afin de pouvoir comparer nos résultats à ceux présentés par les auteurs de cet article pour l'anglais. Cette taxonomie contient deux niveaux [1] : le premier niveau contient cinquante catégories fines, tandis que le second niveau contient six catégories simplifiées. Le tableau 1 présente ces deux niveaux.

2.5. *Corpus de questions*

Pour l'anglais, nous avons utilisé le corpus de questions de (Li *et al.,* 2002), qui provient des collections USC, UIUC et TREC et est en accès libre [2]. Ce corpus a été annoté manuellement selon la hiérarchie de catégories précédente. Les questions d'entraînement sont au nombre de 5 500, celles de test de 500 (ce sont celles de la campagne d'évaluation de questions-réponses TREC 10).

1. http ://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html
2. http ://cogcomp.cs.illinois.edu/Data/QA/QC/

Catégorie simplifiée	Catégorie fine
ABBREVIATION	abbreviation, expression abbreviated
ENTITY	animals, organs of body, colors, inventions...
DESCRIPTION	definition of sth., description of sth., manner of an action, reasons
HUMAN	a group or organization of persons, an individual, title of a person, description of a person
LOCATION	cities, countries, mountains, other locations, states
NUMERIC	postcodes or other codes, number of sth., dates...

Tableau 1. *Classification des questions*

Ces corpus d'entraînement et de test en anglais seront appelés `train_en` et `test_en` par la suite.

Pour le français, nous avons rassemblé les questions de plusieurs campagnes d'évaluation : QA@CLEF 2005, 2006, 2007, EQueR, et Quaero 2008, 2009 et 2010. En ôtant les doublons, nous avons obtenu un corpus de 1 421 questions.

Nous avons annoté manuellement une partie de ce corpus avec les catégories précédentes, et divisé l'ensemble en un corpus d'entraînement de 728 questions, et un corpus de test de 693 questions [3].

Ces corpus d'entraînement et de test en français seront appelés `train_fr` et `test_fr` par la suite.

2.6. *Attributs de classification*

Comme indiqué précédemment, les attributs de base utilisés pour la classification sont les n-grams de mots des questions. Dans les expériences présentées ici, la taille maximale des n-grams est $n = 2$.

Dans un second temps, nous avons annoté les questions avec les catégories morpho-syntaxiques et les lemmes des mots. Nous avons utilisé le TreeTagger (Schmid, 1994) pour effectuer ce traitement, qui a l'avantage de fournir des modèles pour l'anglais et pour le français, ce qui facilite l'adaptation et la comparaison des résultats entre les deux langues.

Enfin, nous avons également utilisé les listes de classes sémantiques provenant de l'analyse des questions par règles de QAVAL. Ces classes correspondent à des listes de mots qui vont déclencher un type de question particulier : ainsi, le mot «président» est dans la classe «Personne» car sa présence dans la question «Quel président français

3. La répartition entraînement/test a été choisie pour être identique à celle utilisée pour la validation de réponses, ce qui explique que les corpus soient de taille quasiment identiques.

a commandé la pyramide du Louvre ?» permet de savoir que cette question attend un nom de personne en réponse.

3. Transfert de la classification

Pour développer le système de classification de questions pour le français, nous ne disposions pas de corpus annoté de grande taille. Nous nous sommes demandé s'il était possible d'utiliser le corpus en anglais. Les travaux sur le transfert de classification et d'annotation sont généralement fondés sur des corpus parallèles, mais en l'absence d'un tel corpus, nous avons voulu étudier si une traduction automatique des questions serait suffisante. Nous avons utilisé l'interface en ligne de Google translate comme système de traduction.

Les méthodes de transfert de classification sont résumées dans la figure 1. Les deux possibilités sont les suivantes :

– dans le premier cas (représenté par des flèches pleines dans la figure), nous traduisons le corpus de questions en anglais vers le français. Nous obtenons ainsi un corpus de questions en français annoté [4], sur lequel il est possible d'apprendre un modèle de classification. Ce modèle est ensuite utilisé sur le corpus de test en français ;

– dans le second cas (représenté par des flèches en pointillés dans la figure), nous apprenons un modèle de classification sur le corpus d'entraînement en anglais. Nous traduisons ensuite les questions du test français vers l'anglais, et appliquons le modèle de classification appris sur la traduction du corpus anglais.

Dans le premier cas, l'avantage est que la classification est apprise sur des questions bien formées ; en revanche, les questions de test étant traduites, les erreurs de traduction pourront perturber le classifieur. Dans le second cas, le modèle de classification sera appris sur un corpus de questions moins bien rédigées mais plus nombreuses que dans le corpus français annoté à la main. Il est donc possible de penser que la plus grande taille du corpus permettra de compenser la perte de qualité.

Afin de comparer les performances à la fois à l'état de l'art et à une classification monolingue, nous avons aussi évalué les performances avec le corpus d'entraînement anglais pour l'apprentissage, et le corpus de test anglais pour le test, et de même pour le français.

4. Expérimentations

Nous avons tout d'abord évalué la classification des questions avec comme attributs les n-grams de mots de la question. Le tableau 2 présente ces premiers résultats.

4. Les systèmes de traduction prenant en entrée des textes bruts, nous séparons les catégories des questions de leurs énoncés, puis les joignons après traduction au nouveau corpus de questions.

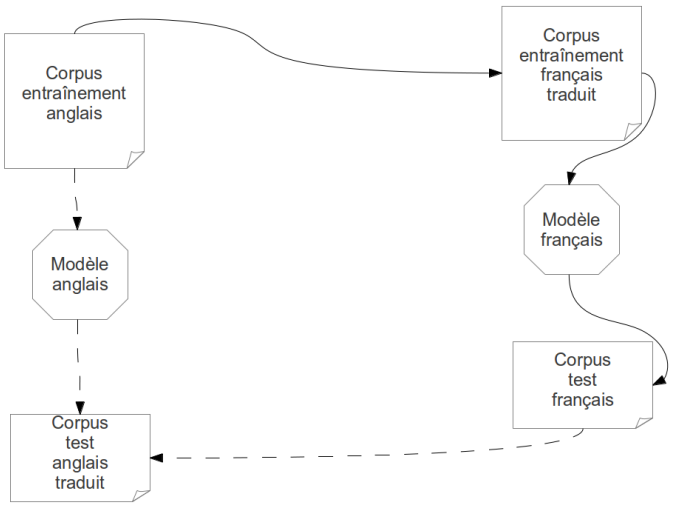

Figure 1. *Deux méthodes de transfert de classification des questions*

Corpus d'entraînement	Corpus de test	Classification fine	Classification simplifiée
anglais	anglais	79,8%	90%
anglais	français traduit en anglais	67,7%	73,5%
anglais traduit en français	français	**79,4%**	82,8%
français	français	76,9%	84%

Tableau 2. *Performance de la classification de questions pour les deux niveaux de hiérarchie (attributs = n-grams de mots, classifieur = libsvm)*

Les performances de classification de question sont données en précision, c'est-à-dire la proportion de questions correctement classées parmi toutes les questions de test.

En se fondant uniquement sur les n-grams de mots, la classification monolingue anglaise obtient donc 79,8% de bonne classification sur la classification fine, et 90% pour la classification simplifiée, résultats qui sont quasiment identiques à ceux trouvés par (Zhang *et al.,* 2003) dans les mêmes conditions pour la classification fine (79,2%), et du même ordre pour la classification simplifiée (87,4%).

Sur le français, nous obtenons des résultats un peu plus bas, de 79,4% pour la classification fine, et de 82,8% pour la classification simplifiée, notamment en raison de la taille plus faible du corpus d'apprentissage : (Zhang *et al.,* 2003) avaient obtenu

Corpus d'entraînement	Corpus de test	Classification fine	Classification simplifiée
anglais	anglais	80,6%	90,6%
anglais traduit en français	français	77,6%	82,4%
français	français	79,4%	85,8%

Tableau 3. *Performance de la classification de questions pour les deux niveaux de hiérarchie (attributs = n-grams de mots, lemmes, et catégories morpho-syntaxiques, classifieur = libsvm)*

une précision de 65% sur la classification fine avec un corpus d'apprentissage réduit à 1 000 questions.

En traduisant les questions du français vers l'anglais, les performances de classification baissent, comme indiqué par (Cumbreras *et al.,* 2006). En revanche, en traduisant le corpus de questions anglais vers le français et en apprenant le modèle de classification sur ce corpus traduit, les performances obtenues sont proches de celles du monolingue français pour la classification simplifiée, et supérieures à celles du monolingue pour la classification fine (et proches du monolingue anglais).

Une explication possible est que la condition où les questions de test sont traduites est très sensible aux erreurs de traduction : si l'une des questions de test est mal traduite, il sera difficile pour le classifieur de la catégoriser correctement. Alors que lorsque le corpus d'entraînement est traduit, les erreurs de traduction peuvent être compensées par d'autres traductions correctes.

Nous avons calculé la significativité des différences de résultats (test de Student, p=0,05) pour chaque niveau de la classification, entre les deux tests en français. Ces différences de résultats sont significatives (ces tests sont faits systématiquement par la suite, et sauf indication du contraire, les différences seront significatives entre deux résultats comparables).

Les questions françaises peuvent donc bien être classifiées par traduction, dans le cas d'une classification fondée sur les n-grams de mots. La traduction du corpus d'entraînement permet d'obtenir des résultats proches des performances monolingues, voire supérieurs. Ces performances restant néanmoins relativement basses en français, surtout pour la classification simplifiée, nous avons ajouté des informations supplémentaires en entrée du classifieur.

Le tableau 3 présente ainsi les performances de classification lorsque le texte en entrée est analysé par l'étiqueteur morpho-syntaxique TreeTagger.

Les résultats diffèrent un peu de ceux n'utilisant que les mots comme attributs, bien que les différences soient faibles. Les performances en monolingue français sont néanmoins meilleures avec cette analyse préalable, mais plus faibles lorsque le corpus

Corpus d'entraînement	Corpus de test	Classification fine	Classification simplifiée
anglais	anglais	82,2%	92%
anglais traduit en français	français	**79,8%**	84,1%
français	français	80,7%	87,2%

Tableau 4. *Performance de la classification de questions pour les deux niveaux de hiérarchie (attributs = n-grams de mots, lemmes, catégories morpho-syntaxiques et classes sémantiques, classifieur = libsvm)*

d'entraînement est le corpus anglais traduit en français. L'étiquetage morphosyntaxique se fondant notamment sur le contexte des mots, il est probable que les mauvaises traductions soient plus difficilement étiquetées.

Enfin, le tableau 4 présente les résultats obtenus en ajoutant les classes sémantiques des mots de la question comme attributs de la classification.

L'ajout des classes sémantiques permet d'améliorer un peu les performances dans toutes les conditions.

5. Travaux similaires

De nombreux systèmes de questions-réponses incluent une classification de questions, qui est généralement fondée sur un apprentissage supervisé. (Li *et al.,* 2002) ont entrainé le classifieur hiérarchique SNoW pour la classification de questions en suivant une hiérarchie fine de 50 classes, et une hiérarchie simplifiée de 6 classes. Ils utilisent comme attributs les mots, les catégories morpho-syntaxiques, les chunks, les entités nommées, les têtes des chunks, et des mots liés à une classe. Leurs meilleurs résultats sont de 98,8% pour la classification simplifiée, et de 95% pour la fine. Cette hiérarchie a été largement utilisée par les autres systèmes de questions-réponses par la suite.

(Zhang *et al.,* 2003) notamment ont étudié l'apport des arbres syntaxiques des questions pour leur classification. Ils obtiennent leurs meilleurs résultats avec le corpus d'apprentissage le plus grand, les SVM et des n-grams de mots comme attributs dans le cas d'attributs vectoriels (87% de bonne classification pour les classes simplifiées, et environ 80% pour les classes fines), et avec les arbres syntaxiques de constituants dans le cas d'attributs structurés et des tree kernels (90% pour les classes simplifiées, et 80% également pour les classes fines).

Ces classifications des questions été appliquées uniquement à l'anglais. Bien entendu, ces méthodes peuvent être adaptées à d'autres langues, mais cela nécessite

l'annotation de grands corpus de questions, et des outils (notamment un analyseur syntaxique) disponibles dans chacune des langues.

Afin de développer une classification multilingue sans avoir à développer de nouveaux outils pour toutes les langues, (Solorio *et al.,* 2004) ont proposé une approche consistant à exploiter internet pour déterminer le type attendu. En combinant les informations obtenues par requêtes avec les mots de la question, ils obtiennent 84% de bonne classification pour l'anglais, 84% pour l'espagnol, et 89% pour l'italien, avec une cross-validation sur un corpus de 450 questions, pour 7 catégories de questions. L'une des limitations évoquée par les auteurs est le manque de corpus annoté de grande taille disponible pour toutes les langues.

Nous nous plaçons dans ce même cas et essayons d'étudier la possibilité d'utiliser les ressources créées pour l'anglais. (Cumbreras *et al.,* 2006) se sont intéressés à la même problématique, afin de développer un système pour l'espagnol. Ils utilisent deux outils de traduction en ligne pour traduire leurs questions de l'espagnol vers l'anglais, puis un système de classification de questions pour l'anglais. Ils obtiennent une précision de 65% sur la classification des questions espagnoles, alors que les questions anglaises sont bien classifiées à 80%. Cette méthode mène donc à une baisse de performance assez importante.

Le problème de la disponibilité de corpus d'apprentissage se pose également dans le cas de systèmes de questions-réponses crosslingues, pour lesquels la question est dans une langue, et les documents dans lesquels chercher la réponse dans une autre. La stratégie des systèmes est généralement de traiter autant que possible de l'anglais, par exemple en traduisant les réponses uniquement (Bos *et al.,* 2006, Bowden *et al.,* 2008).

6. Conclusion

Nous nous sommes intéressés à la classification de questions pour un système de questions-réponses en français. Puisqu'il n'existe pas de corpus de questions en français annoté pour cette tâche, nous avons souhaité exploiter un corpus existant de grande taille en anglais, et avons testé deux façons d'utiliser ce corpus : par traduction du corpus d'entraînement, et par traduction du corpus de test. Nous avons mené des expériences sous différentes conditions, et avons montré qu'il est possible d'obtenir une classification de questions pour le français en traduisant le corpus d'entraînement.

Les annotations morpho-syntaxiques n'améliorant pas la qualité de la classification, nous n'avons pas testé l'utilisation des arbres syntaxiques. L'utilisation de cette information donne cependant les meilleurs résultats sur l'anglais (Zhang *et al.,* 2003), donc il pourrait néanmoins être intéressant de mener cette étude.

Par ailleurs, les formes interrogatives étant peu présentes dans les corpus utilisés par les systèmes de traduction, elles sont généralement moins bien traduites. Une perspective de ce travail serait d'utiliser un modèle de traduction spécifique pour les ques-

tions afin de pouvoir apprendre la classification sur des traductions de meilleure qualité.

7. Bibliographie

Bos J., Nissim M., « Cross-lingual question answering by answer translation », *Working Notes of the Cross Language Evaluation Forum*, 2006.

Bowden M., Olteanu M., Suriyentrakorn P., d´Silva T., Moldovan D., « Multilingual question answering through intermediate translation : Lcc´s poweranswer at qa@clef 2007 », *Advances in Multilingual and Multimodal Information Retrieval*, vol. 5152, p. 273-283, 2008.

Chang C.-C., Lin C.-J., « LIBSVM : A library for support vector machines », *ACM Transactions on Intelligent Systems and Technology*, vol. 2, p. 27 :1-27 :27, 2011.

Cumbreras M., López L., Santiago F., « Bruja : Question classification for spanish. using machine translation and an english classifier », *Proceedings of the Workshop on Multilingual Question Answering*, Association for Computational Linguistics, p. 39-44, 2006.

Grappy A., Grau B., Falco M.-H., Ligozat A.-L., Robba I., Vilnat A., « Selecting answers to questions from Web documents by a robust validation process », *IEEE/WIC/ACM International Conference on Web Intelligence*, 2011.

Jabaian B., Besacier L., Lefèvre F., « Combination of stochastic understanding and machine translation systems for language portability of dialogue systems », *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, p. 5612-5615, 2011.

Li X., Roth D., « Learning question classifiers », *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, p. 1-7, 2002.

Ligozat A.-L., Exploitation et fusion de connaissances locales pour la recherche d'informations précises, PhD thesis, Université Paris-Sud, 2006.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49, 1994.

Solorio T., Pérez-Coutino M. et al., « A language independent method for question classification », *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, p. 1374-1380, 2004.

Zhang D., Lee W., « Question classification using support vector machines », *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ACM, p. 26-32, 2003.