Classification des questions d'opinion dans un système de Questions-Réponses pour les débats politiques

Amine Bayoudhi [1]

Équipe ANLP – Laboratoire MIRACL
FSEG, Université de Sfax,
B.P. 1088, 3018, Sfax TUNISIE
bayoudhi.amine@gmail.com

RÉSUMÉ. *L'identification du type de la question est une étape importante dans le développement des Systèmes de Questions-Réponses (SQR). Ramenée le plus souvent à un problème de classification, cette étape vise à déterminer le type de la réponse attendue à la question en attribuant à la question une ou plusieurs classes selon la taxonomie adoptée. Dans la littérature, la plupart des SQR se sont orientés vers des sources d'information monologales. Les conversations, bien qu'elles constituent une partie importante des sources d'information, manquent d'outils robustes de recherche d'information tels que les SQR. Dans cet article, nous proposons une nouvelle taxonomie et une approche pour la classification des questions d'opinion dans le cadre d'un SQR arabe pour les débats politiques. Les résultats obtenus sont pertinents et atteignent une précision moyenne de l'ordre de 89,51 %.*

ABSTRACT. *The identification of the question type is an important task in Question Answering Systems (QAS). Most commonly considered as a classification problem, this task aims to determine the type of the expected answer by assigning to the question one or more classes according to the adopted taxonomy. Most QAS were focusing on monologal sources. Conversations, although they constitute important information sources, lack robust tools for information retrieval such as QAS. In this paper, we propose a new taxonomy and a method to classify opinion questions as part of an Arab QAS for political debates. We obtained relevant results with a mean precision of around 89,51%.*

MOTS-CLÉS : *taxonomie des types de questions, classification des questions, extraction d'opinion, système de Questions-Réponses.*

KEYWORDS: *question types taxonomy, questions classification, opinion extraction, Question Answering systems.*

1. Directeurs de thèse :
Hatem Ghorbel, Hatem.Ghorbel@he-arc.ch
Haute École Arc Ingénierie, Rue de la serre 7, CH-2610 St-Imier SUISSE.
Lamia Hadrich Belguith, L.Belguith@fsegs.rnu.tn
ANLP – Laboratoire MIRACL, FSEG, Université de Sfax, B.P. 1088, 3018, Sfax TUNISIE.

1. Introduction

Avec l'énorme quantité des archives d'information, trouver l'information la plus appropriée avec le minimum d'effort constitue de nos jours un besoin quotidien. Les systèmes de Questions-Réponses (SQR) répondent à ce besoin. En effet, ils offrent à l'utilisateur la possibilité de formuler ses requêtes en langage naturel, et lui fournissent des réponses précises et restreintes. Les SQR traitent aussi bien les questions factuelles que celles non factuelles. Dans la littérature, les questions factuelles ont fait l'objet de beaucoup de travaux de recherche. Cependant, les questions non factuelles, en particulier celles d'opinion, n'ont pas eu le même sort car elles sont plus complexes et requièrent des analyses assez profondes.

L'identification du type de la question est une étape importante dans le développement des SQR. Ramenée le plus souvent à un problème de classification, cette étape vise à déterminer le type de la réponse attendue à la question en attribuant à la question une ou plusieurs classes selon la taxonomie adoptée.

Le présent travail se situe dans le contexte du développement d'un SQR arabe pour les débats politiques. Ainsi, nous présentons dans la section 2 quelques travaux similaires sur la classification automatique des questions. Dans les sections 3 et 4, nous proposons respectivement notre nouvelle taxonomie et notre approche pour la classification des questions d'opinion. Dans la section 5, nous terminons par les conclusions et les perspectives.

2. Travaux similaires

Nous distinguons dans la littérature trois approches de classification des questions :

– *L'approche à base de règles* consiste à associer à la question un ensemble de règles définies manuellement et appelées *hand-crafted rules* (Prager *et al.*, 1999). Elle se base généralement sur les articles interrogatifs utilisés dans les questions (Kwok *et al.*, 2001).

– *L'approche à base d'apprentissage* consiste à extraire un ensemble de critères à partir de la question et à construire un modèle de classification afin d'inférer le type adéquat de la question. Les travaux adoptant cette approche diffèrent *i)* selon les types de classificateurs utilisés tels que les réseaux de neurones (Loni, 2011) et les arbres de décision (Tomuro, 2002) *ii)* selon les critères d'apprentissage choisis qui peuvent être *symboliques* (Razmara *et al.*, 2007), morpho-syntaxiques (Khoury, 2011), sémantiques (Cai *et al.*, 2006) ou statistiques (Ku *et al.*, 2008).

– *L'approche hybride* consiste à combiner les deux approches précédentes en utilisant comme critères d'apprentissage des règles prédéfinies manuellement (Silva *et al.*, 2011).

3. Taxonomie proposée pour la classification des questions

Pour repérer les types de questions dans un SQR pour les débats politiques, nous avons construit le corpus d'étude *COPARQ* (Corpus of OPinion ARabic Questions) dont les caractéristiques sont illustrées par le tableau 1.

Caractéristiques du corpus	Valeurs
Nombre de discussions	14
Nombre de rédacteurs	14
Taille totale en nombre de questions	620
Taille totale en nombre de mots	7 549

Tableau 1. *Caractéristiques du corpus d'étude COPARQ*

Après l'étude du corpus, nous avons conçu une taxonomie pour les questions d'opinion dans le cadre d'un SQR arabe pour les débats politiques. Cette taxonomie est basée sur la nature de l'information sur laquelle porte la question et sur la nature de la réponse attendue. Elle est inspirée du modèle proposé par Paroubek et al. pour l'évaluation du projet DOXA (Paroubek *et al.*, 2010).

3.1. *Catégories des questions*

Nous proposons dans un premier temps 3 catégories principales de classification de questions d'opinion dans le cadre d'un SQR arabe pour les débats politiques :

– *Thematic* : c'est la catégorie des questions où on se demande si un locuteur est intervenu dans la discussion d'un thème donné, ou simplement de reporter l'intervention d'un locuteur dans un thème.

– *Informational* : c'est la catégorie des questions dans lesquelles l'aspect factuel est dominant sur l'aspect d'opinion. Elle contient généralement les questions portant sur une information ou un évènement/concept/personne selon l'opinion d'un locuteur donné.

– *Opinionated* : c'est la catégorie des questions portant sur les attributs principaux de l'opinion tels que l'attitude, le porteur de l'opinion (*Source*) et la cible (*Target*) envers lesquels l'attitude est exprimée.

3.2. *Classes d'opinion*

Etant donné que notre SQR est destiné aux questions d'opinion, nous nous intéressons dans le présent travail à la catégorie *Opinionated*. C'est pourquoi, nous avons procédé dans un deuxième temps à un niveau supplémentaire de classification de la catégorie *Opinionated* en s'inspirant de la classification de Ku et al. (Ku *et al.*, 2008). Ainsi, pour cette catégorie nous définissons sept classes (Figure 1) :

– *Attitude* : demander l'attitude du porteur de l'opinion envers la cible donnée.

– *Yes/No* : demander si le porteur d'opinion adopte l'attitude donnée.

– *Holder* : demander qui a exprimé l'attitude donnée envers la cible donnée.

– *Target* : demander envers qui ou quoi le porteur d'opinion donné a exprimé l'attitude donnée.

– *Reason* : demander les raisons pour lesquelles le porteur d'opinion a adopté l'attitude donnée envers la cible donnée.

– *Majority* : demander laquelle des options est majoritaire.

– *Intensity* : demander à quel point le porteur de l'opinion donné adopte l'attitude donnée envers la cible donnée.

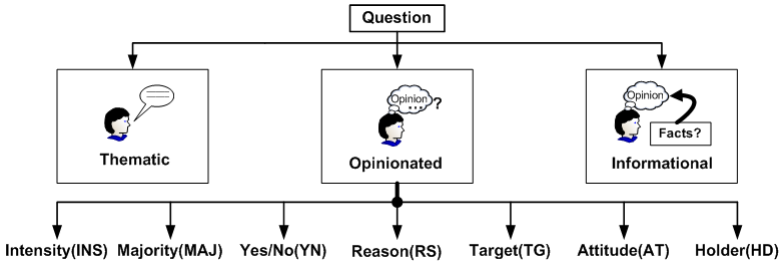

Figure 1. *Taxonomie proposée pour la classification des questions dans un SQR pour les conversations*

4. Méthode proposée pour la classification des questions d'opinion

Notre approche de classification des questions s'inspire des techniques d'extraction de l'opinion et se base sur les techniques d'apprentissage supervisé. Elle est constituée de deux phases principales : l'extraction des critères d'apprentissage et la classification automatique des questions (Figure 2).

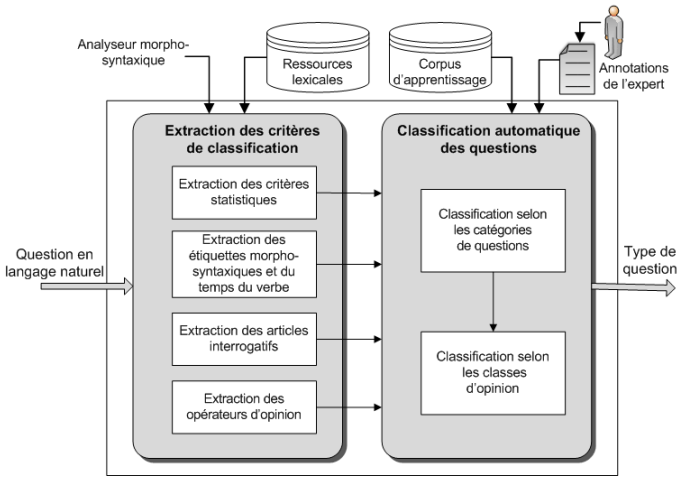

Figure 2. *Méthode proposée pour la classification des questions d'opinion*

La phase de l'extraction des critères de classification est effectuée en quatre étapes:

– *Extraction des étiquettes morphosyntaxiques et du temps des verbes* : permet d'extraire les critères morphosyntaxiques en utilisant un analyseur morpho-syntaxique. Cette étape permet aussi de détecter le temps du verbe si la question contient un groupe nominal verbal.

– *Extraction des articles interrogatifs* : permet d'extraire les articles interrogatifs en utilisant des listes exhaustives d'articles interrogatifs tel que « من » (qui), d'articles interrogatifs attachés à des prépositions tel que « لأيّ » (pour quel), et de verbes à l'impératif utilisés dans le contexte interrogatif tel que « اذكر » (cite).

– *Extraction des marqueurs d'opinion* : permet d'extraire les marqueurs d'opinion existant dans la question en utilisant des listes de marqueurs comme les verbes d'opinion, les noms, les adjectifs ou les adverbes tels que « اعتقد » (croire), « رأي » (avis), « إيجابي » (positif) et « أفضل » (meilleur).

– *Extraction des critères statistiques* : permet d'extraire les critères statistiques en se référant au nombre de mots dans la question. Elle permet aussi de calculer les probabilités des *unigrammes* et des *bigrammes* tels que « أكد » (confirmer), « علق » (commenter), « حول موضوع » (à propos du sujet de). Ces *unigrammes* et ces *bigrammes* sont utilisés surtout pour identifier la catégorie *Thematic*.

Dans le processus de classification, nous avons appliqué l'algorithme SVM sous l'environnement *Weka* [2] sur un corpus d'entrainement collecté de plusieurs sources. Les caractéristiques de ce corpus sont illustrées par le tableau 2.

Corpus	**Nombre total de questions**	**Taille totale (en mots)**	**Taille moyenne par question (mots)**
COPARQ	620	7 531	12,146
Conférences	723	6 000	8,298
Polls	596	5 915	9,942
Total	1 939	19 446	10,028

Tableau 2. *Caractéristiques du corpus d'entrainement*

Nous avons évalué notre approche de classification en termes de précision [1] qui mesure la capacité à bien classer la question dans la classe adéquate.

$$Précision = \frac{nombre\ de\ questions\ \ bien\ \ classées}{nombre\ \ total\ \ de\ questions} \qquad [1]$$

Cette mesure de précision est calculée après l'application de la méthode d'évaluation validation croisée en dix sous ensembles. Les résultats de la classification selon les catégories de questions et selon les classes d'opinion ont atteint respectivement 87,9 % et 91,13 %. La précision moyenne est de **89,51 %**.

2. http ://www.cs.waikato.ac.nz/ ml/weka/index.html

5. Conclusions et perspectives

Dans cet article, nous avons défini une taxonomie pour la classification des questions dans un SQR d'opinion pour les débats politiques, en s'inspirant des modèles de fouille d'opinion et d'analyse de sentiments. De plus, nous avons proposé une approche pour la classification des questions d'opinion basée sur des critères de surface. Les résultats ont été bons et atteignent 89,51 % comme mesure de précision moyenne.

Comme perspectives, nous envisageons de construire un dictionnaire de sentiments pour collecter les marqueurs d'opinion et leurs assigner des degrés de subjectivité. Le dictionnaire nous permettra de mieux détecter la nature subjective des questions et par la suite d'améliorer les résultats obtenus dans la classification des catégories de questions.

Bibliographie

Cai D., Sun J., Zhang G., Lv D., Dong Y.; Song Y.; Yu C., « HowNet Based Chinese Question Classification », *Pacific Asia Conference on Language, Information and Computation PACLIC'06*, Wuhan, 1-3 Novembre 2006, p. 366-369.

Khoury R., « Question Type Classification Using a Part-of-Speech Hierarchy », *Autonomous and Intelligent Systems AIS'11*, Burnaby, 22-24 Juin 2011, Springer, p. 212-221.

Ku L.W., Liang Y.T., Chen H.H.: « Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems ». *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 13, n° 3, 2008, p. 307-326.

Kwok C., Etzioni O., Weld D.S., « Scaling question answering to the web », *international conference on World Wide Web WWW'01*, Hong Kong, 1-5 Mai 2001, p. 150-161.

Loni B., « Enhanced Question Classification with Optimal Combination of Features », Master of Science Thesis, Delft University of Technology, 2011.

Paroubek P., Pak A., Mostefa D., « Annotations for Opinion Mining Evaluation in the Industrial Context of the DOXA project ». *international conference on Language Resources and Evaluation LREC'2010*, Malta, 17-23 Mai 2010, p. 1105-1112.

Prager J., Radev D., Brown E., Coden A., « The use of predictive annotation for question answering in trec8 ». *Text REtrieval Conference NIST TREC 8*, Gaithersburg, 17-19 Novembre 1999, p. 399-411.

Razmara M., Fee A., Kosseim L., « Concordia University at the TREC 2007 QA track ». *Text REtrieval Conference TREC'2007*, Gaithersburg, 5-9 Novembre 2007.

Silva J., Coheur L., Mendes A.C., Wichert A., « From symbolic to sub-symbolic information in question classification », *Artificial Intelligence Review*, Vol. 35, 2011, p. 137-154.

Tomuro, N., « Question terminology and representation of question type classification ». *International Workshop on Computational Terminology COMPUTERM'*2002, Taipei, 24 Août-1 Septembre 2002.