Réponse à des tests de compréhension

Brigitte Grau — Anne-Laure Ligozat — Van-Minh Pho

LIMSI-CNRS
rue John von Neumann

91403 Orsay Cedex
prenom.nom@limsi.fr

RÉSUMÉ. *Dans cet article, nous présentons une adaptation d'un système de questions-réponses existant pour une tâche de réponse à des questions de compréhension de textes. La méthode proposée pour sélectionner les réponses correctes repose sur la reconnaissance d'implication textuelle entre les hypothèses et les textes. Les spécificités de cette méthode sont la génération d'hypothèses par réécriture syntaxique, et l'évaluation de plusieurs critères de distance, adaptés pour gérer des variantes de termes.*

ABSTRACT. *This paper presents an adaptation of an existing question answering system for a machine reading task. The proposed method for selecting correct answers relies on textual entailment recognition between hypotheses and texts. The specificities of this method are the generation of hypotheses by parse tree modification and the evaluation of several distance criteria, adapted to take into account term variants.*

MOTS-CLÉS : *Systèmes de questions-réponses, implication textuelle, compréhension de textes.*

KEYWORDS: *Question answering systems, textual entailment, text comprehension.*

1. Introduction

L'évaluation du niveau d'acquisition d'une langue ou d'un domaine est souvent faite au moyen de QCM dont les questions sont posées sur un texte de référence permettant de trouver les réponses. Répondre à des questions de compréhension de textes repose souvent sur l'utilisation de connaissances structurées sur le domaine étudié, et de mécanismes d'inférences adaptés (Clark *et al.,* 2012). L'un des processus sous-tendant ces inférences est l'implication textuelle (*textual entailment*), qui vise à décider si un énoncé en implique un autre, et la reconnaissance d'implication textuelle est généralement utilisée comme élément de base dans les processus de validation de réponse (Peñas *et al.,* 2007, Penas *et al.,* 2008, Rodrigo *et al.,* 2009). La validation de réponse comporte généralement deux sortes de méthodes : une vérification du type de la réponse (Grappy *et al.,* 2011) , et une reconnaissance d'implication textuelle. En ce qui concerne la reconnaissance d'implication textuelle, les méthodes généralement utilisées sont fondées sur un apprentissage supervisé exploitant des attributs liés au recouvrement lexical, à la densité des termes communs et à des structures syntaxiques analogues (Dagan *et al.,* 2006).

Nous nous sommes intéressés à la problématique de réponse à des QCM dans le cadre de la campagne d'évaluation QA4MRE, dont l'édition 2012 proposait l'évaluation de cette tâche sur la maladie d'Alzheimer. L'exemple suivant montre une question, les cinq réponses possibles, ainsi que la phrase justificative de la bonne réponse.

> Question : Which enzyme is responsible for the transformation of testosterone into estrogen ?
>
> R1) aromatase (bonne réponse) R4) BACE1
>
> R2) AD R3) androgen R5) actinomycin D
>
> Document : As brain testosterone plays both androgenic and estrogenic actions due to its conversion into estrogen via aromatase naturally, it is unclear that the age-related reduction of testosterone increased risk of Alzheimer's disease (AD) in men is mediated through androgen alone or both androgen and estrogen mechanisms.

Les systèmes participant à QA4MRE commencent généralement par effectuer une analyse des questions pour générer des hypothèses, principalement par patrons, écrits manuellement ou acquis. La plupart des systèmes utilisent ensuite des pré-traitements, tels que la résolution d'anaphores ou de coréférence, ou la reconnaissance d'entités nommées. Le système ayant obtenu les meilleurs résultats à la tâche générale de QA4MRE 2011 et 2012 (Pakray *et al.,* 2011, Bhaskar *et al.,* 2012) utilise plusieurs mesures d'implication textuelle (comparaison des entités nommées, n-grammes et skip n-grammes communs), et compare le type de réponse avec celui attendu. En parallèle, ils utilisent un système de questions-réponses, et comparent ensuite les scores obtenus par les deux systèmes.

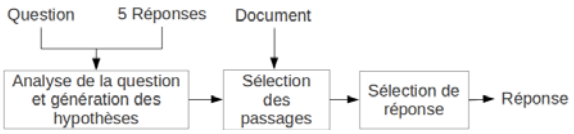

Figure 1. *Architecture du système QALC adapté à QA4MRE*

Notre approche a consisté à partir du système de questions-réponses existant, QALC (de Chalendar *et al.,* 2002), que nous avons adapté à cette tâche en utilisant une reconnaissance d'implication textuelle pour la validation de réponses en réécrivant la question sous forme déclarative avec chacune des réponses, telle que «Aromatase is responsible for the transformation of testosterone into estrogen» pour l'exemple précédent. Notre approche se distingue de celles d'autres systèmes de réponse à des questions de compréhension sur deux points principaux : nous avons développé un module de réécriture syntaxique de la question, afin de pouvoir appliquer des critères de distance sur des représentations syntaxiques ; par ailleurs, nous avons utilisé des ressources spécifiques pour trouver des variations, notamment pour le domaine biomédical, et adapté les distances existantes aux variations trouvées. Cette approche nous situe 3ème sur le corpus d'évaluation de QA4MRE, avec une précision de 0,395.

2. Système de réponse à des QCM

Nous avons réutilisé notre système de questions-réponses QALC pour l'anglais, en adaptant l'architecture (voir figure 1). Par rapport au système d'origine, le système utilisé pour QA4MRE prend en entrée les réponses possibles en plus de la question. Une étape de génération des hypothèses est ajoutée à l'analyse de la question. Enfin, après la sélection de passages, un module sélectionne la réponse après classement, au lieu d'extraire les réponses des passages. Nous détaillons par la suite les modifications que nous y avons effectuées.

2.1. *Analyse des questions*

L'objectif du module d'analyse des questions est double :

– déterminer le type de réponse attendu, comme «enzyme» dans l'exemple donné.

– reformuler la question à l'affirmative pour former les hypothèses, comme «Aromatase is responsible for the transformation of testosterone into estrogen».

Ce module a été développé pour QA4MRE car les types de questions et réponses sont très différents en domaine biomédical par rapport au domaine général pour lequel QALC avait été conçu. Il s'appuie sur l'analyse syntaxique de la question par Biomodel (McClosky, 2010), et sur un module de génération existant (Pho, 2012).

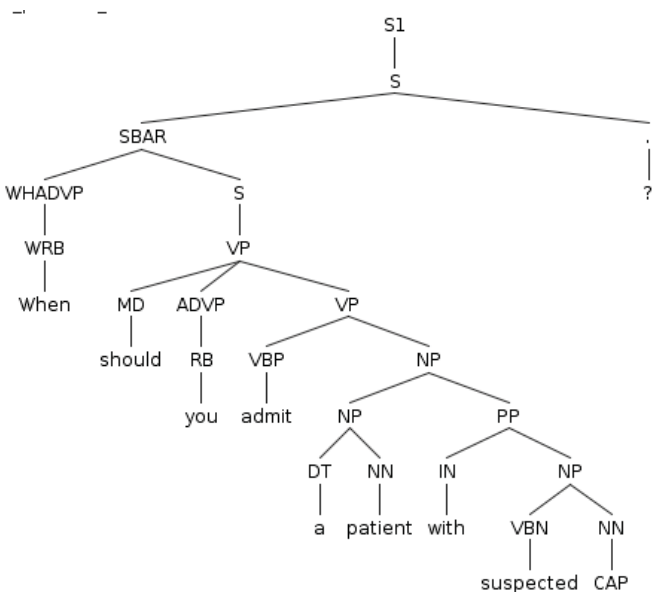

Figure 2. *Arbre syntaxique d'une question utilisé pour produire l'hypothèse*

La détermination du type attendu s'appuie sur la position du type dans l'arbre syntaxique de la question par rapport à la position de l'interrogatif. Tregex (Levy *et al.,* 2006) permet de reconnaître le nœud répondant aux critères voulus dans l'arbre.

L'arbre de constituants de la question est ensuite utilisé pour générer les hypothèses correspondant à chaque réponse proposée. Des règles de manipulation d'arbre par Tsurgeon (Levy *et al.,* 2006) permettent l'inversion éventuelle de constituants et le remplacement de l'interrogatif par chaque réponse possible. La sortie contient donc cinq hypothèses contenant chacune une réponse possible. Ainsi, pour l'arbre de la figure 2, les hypothèses générées seront les arbres de constituants correspondant au modèle «You should admit a patient with suspected CAP when <REPONSE>».

Les règles de détection du type et de reformulation ont été développées sur un corpus de 300 questions médicales extraites du Journal of Family Practice[1]. La détermination du type attendu est correcte dans 94% des cas sur ce corpus de questions, tandis que la génération d'hypothèses fournit des hypothèses correctes pour 74% des questions. Les 6% d'erreurs de détection du type proviennent essentiellement d'analyses syntaxiques erronées : par exemple, la question «Which medications can be split without compromising efficacy and safety ?», le type attendu est censé être «medications» mais l'analyseur ne trouve pas de type attendu car l'analyseur syntaxique

1. www.jfponline.com

a annoté «medications» comme un verbe. Concernant la génération d'hypothèses, les erreurs viennent soit de l'analyse syntaxique, soit de formes de questions non prévues.

2.2. *Sélection des passages*

Les passages sont formés de chaque phrase du document support. Nous avons normalisé les lettres grecques, afin d'unifier leur écriture entre les questions, les réponses et les documents. La sélection de phrase est ensuite fondée sur la reconnaissance des termes de l'hypothèse ou de leurs variantes. L'outil Fastr (Jacquemin, 1999) est utilisé pour reconnaître des variantes morphologiques, syntaxiques et sémantiques de termes, en utilisant des lexiques de variations morphologiques et sémantiques. Nous avons complété les variantes sémantiques de Fastr issus de WordNet par les variantes contenues dans FrameNet (Baker *et al.,* 1998). Nous avons ainsi extrait 673 entrées, avec plusieurs variantes pour chaque entrée : par exemple, le nom «modification» a comme variantes le verbe «change», le verbe «convert»... Ces entrées contiennent plusieurs variantes utiles pour notre tâche, comme «convert»/«transformation», mais aussi des antonymes comme «successful»/«failed».

Nous attribuons un poids, nommé `PQALC`, à chacune des phrases selon la mesure mise au point pour QALC (de Chalendar *et al.,* 2002) qui repose sur les attributs suivants :

– les mots de la question, pondérés par leur degré de spécificité,

– les variantes des mots de la question,

– les mots exacts de la question,

– la proximité des mots de la question entre eux.

Le poids de la phrase est d'abord calculé en fonction de la présence des mots de la question dans la phrase, puis les poids issus des autres critères sont ajoutés.

Le poids de base de la phrase est calculé à partir des lemmes des mots (ou des mots si le mot est inconnu de l'étiqueteur), et de leur degré de spécificité. Celui-ci est calculé à partir de l'inverse de la fréquence du mot dans un corpus journalistique de grande taille. Dans le cas de la maladie d'Alzheimer, de nombreux termes sont spécifiques et absents de ce corpus, ce qui leur confère une spécificité de 1, valeur maximale. Certains mots ne sont pas pris en compte : déterminants, prépositions, auxiliaires et noms transparents. Un nom transparent est un nom dont le complément est plus pertinent que le mot lui-même. Par exemple, le mot «kind» est transparent dans une question comme «What kind of gliar cell...», et «cell» est le mot sémantiquement pertinent. Nous avons constitué une liste de ces mots. Le poids de base d'une phrase est donné par : $BasicWeight = \frac{\sum_{i=1}^{m} dr_i}{\sum_{j=1}^{n} dr_j}$ avec :

– dr_i le poids d'un terme de la question trouvé dans la phrase. Si ce terme est un mono-terme (formé d'un seul mot), le poids est le degré de spécificité du mot, sinon le poids est fixé à 0,1 ;

– dr_j le poids d'un terme de la question. Si ce terme est un mono-terme (formé d'un seul mot), le poids est le degré du mot, sinon le poids est fixé à 0,1 ;

– m nombre de termes trouvés dans la phrase,

– n nombre de termes dans la question.

Chaque terme n'est compté qu'une seule fois dans la formule, même s'il est trouvé plusieurs fois dans la phrase. Si un mot de la question n'est pas trouvé à l'identique dans la phrase mais sous forme d'une variante, son poids est divisé par deux. Comme les poids des mots appartiennent à l'intervalle [0-1], le maximum du poids d'une phrase est proche de 1. Pour qu'il soit plus facilement lisible, nous le multiplions par 1 000. Nous ajoutons ensuite des poids supplémentaires à ce poids de base pour chaque critère supplémentaire satisfait. Chaque poids ajouté ne peut pas être supérieur à 10% du poids de base. Le critère de proximité des mots de la question entre eux cherche à représenter le fait que plusieurs mots sont utilisés de la même façon dans la question et la phrase. Il est calculé pour les unitermes deux à deux. Chaque paire de mots qui est séparée au maximum d'un mot significatif reçoit un poids de 0,02. Le dernier critère représente le nombre de lemmes trouvés dans la phrase sans variation. Au final, le poids $PQALC$ d'une phrase S est :

$$PQALC(S) = BasicWeight \times 1000 + MutualCloseness \times 1000 + ExactLemmas \times 100$$

Les réponses candidates sont également recherchées dans la phrase, et sont pondérées selon les mêmes critères, sans le critère de proximité.

$$P\_REP(a) = BasicWeight \times 1000 + ExactLemmas \times 100$$

2.3. *Pondération des réponses*

Afin de pouvoir sélectionner la bonne réponse parmi les réponses candidates, outre la mesure `P_REP`, présentée en 2.2, nous évaluons celles-ci selon la proximité de l'hypothèse qui leur est associée avec les phrases sélectionnées.

Ainsi, pour l'exemple, «aromatase» devrait être sélectionnée car «aromatase is responsible for the transformation of testosterone into estrogen» est proche de «As brain testosterone plays both androgenic and estogenic actions due to its conversion into estrogen via aromatase naturally (..)». La difficulté est de déterminer et combiner les critères à utiliser pour trouver que ces phrases sont sémantiquement proches : ici, l'analyse syntaxique permettrait de relier «testosterone» à «conversion» et au reste des mots de l'hypothèse dont la réponse (car «its» l'a pour antécédent), et il faudrait aussi reconnaître que «conversion» est synonyme du mot de l'hypothèse «transformation».

Afin de déterminer les distances entre hypothèses et phrases des documents, nous avons choisi d'utiliser des métriques du domaine de l'implication textuelle qui exploitent deux niveaux de représentation des textes : le niveau surfacique et le niveau syntaxique. Nous avons calculé plusieurs mesures de similarité, fondées sur les caractéristiques communes, les distances d'édition et les alignements. Nous les avons appliquées en tenant compte des variantes trouvées.

Nous avons calculé des mesures de recouvrement de sous-phrases entre les représentations syntaxiques, sans tenir compte de leur position dans la phrase. L'une est fondée sur le nombre de dépendances communes, donc de liens communs, entre hypothèse et texte, pouvant associer une dépendance à un chemin de dépendances entre termes équivalents. La seconde mesure, reprise de (Wang *et al.,* 2009) est fondée sur le nombre de sous-arbres communs à l'hypothèse et au texte calculé avec une fonction de «*tree kernel*» entre les arbres syntaxiques de ceux-ci. Ces deux mesures n'ayant pas donné les meilleurs résultats, nous ne les détaillons pas ici.

2.3.1. *Distance d'édition sur les arbres*

La mesure `treeEdit` prend en compte la représentation complète des phrases et calcule le nombre d'opérations minimal transformant l'arbre syntaxique du texte en celui de l'hypothèse, soit le nombre d'insertions, de substitutions et de suppressions de nœuds. Nous l'avons calculée sur les arbres de dépendances du texte et de l'hypothèse et avons implémenté l'algorithme de Zhang et al. (Zhang *et al.,* 1989).

2.3.2. *Alignement des formes de surface*

TERp (Translation Edit Rate plus) (Snover *et al.,* 2009) est un algorithme calculant la distance d'édition pour transformer le texte en l'hypothèse sur les suites de termes. La différence avec une distance d'édition classique est qu'elle aligne les termes équivalents de l'hypothèse et du texte et déplace les parties du texte afin de minimiser la distance d'édition. Dans notre cas, nous ne considérons pas le texte dans son ensemble : seul le fragment contenant les mots communs à l'hypothèse nous intéresse.

2.4. *Sélection des réponses*

Une fois que toutes les réponses ont été évaluées selon les différents critères, il faut sélectionner la réponse qui semble la plus pertinente. Plusieurs modes de sélection ont été définis, après ordonnancement des phrases selon le poids `PQALC`. Les poids des réponses possibles sont donc `P_REP`, `TERp` ou `treeEdit` :

– la réponse la plus présente dans les *n* premières phrases retournées par le système. Après expérimentations, nous avons fixé *n* à 5. En cas d'égalité de fréquence, la réponse dans la phrase de meilleur poids est sélectionnée, et s'il y a plusieurs réponses dans la même phrase, la réponse de meilleur poids est sélectionnée. Ce mode de sélection sera noté `freqTop` par la suite ;

	Alzheimer			Domaine général		
	`freq`	`freqTop`	`maxSTop`	`freq`	`freqTop`	`maxSTop`
`P_REP`	10 / 0,25	14 / 0,367	-	60 / 0,382	56 / 0,370	**62 / 0,395**
`TERp`	11 / 0,288	**15 / 0,393**	9 / 0,225	50 / 0,318	50 / 0,330	51 / 0,325
`treeEdit`	12 / 0,315	**15 / 0,393**	14 / 0,35	52 / 0,331	52 / 0,343	53 / 0,337
baseline	8 / 0,2			32 / 0,2		

Tableau 1. *Résultats sur les deux corpus en nombre de bonnes réponses sélectionnées*

– la réponse la plus présente dans les *n* premières phrases retournées par le système contenant une réponse (*n*=5), avec les mêmes options en cas d'égalité. Ce mode de sélection sera noté `freq` par la suite ;

– la réponse de poids maximal dans les *n* premières phrases retournées par le système. Ce mode de sélection sera noté `maxSTop` par la suite.

3. Expérimentations

3.1. *Corpus et mesure d'évaluation*

Pour évaluer notre système, nous avons utilisé deux jeux d'évaluation provenant des campagnes d'évaluation QA4MRE pour l'anglais, l'un pour le domaine général, comportant quatre domaines, et l'autre sur la maladie d'Alzheimer, notre tâche principale. Dans les deux cas, étaient fournis, par domaine, une collection de documents du même type pouvant servir de source de connaissances, quatre documents sur lesquels étaient posées dix questions de compréhension, et pour chaque question, cinq choix de réponses. La mesure d'évaluation utilisée dans QA4MRE est appelée *c@1* :

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n})$$

avec n le nombre total de questions, n_U le nombre de questions auxquelles le système n'a pas répondu, et n_R le nombre de questions auxquelles le système a correctement répondu. Nous présentons également nos résultats sous la forme plus simple du nombre de bonnes réponses retournées. La baseline considérée dans l'évaluation correspond à un système qui choisirait une réponse au hasard parmi les cinq choix.

3.2. *Résultats*

Pour chaque tâche, nous ne présentons qu'une sélection des meilleurs résultats. Le tableau 1 présente les résultats obtenus dans le domaine d'Alzheimer et dans le domaine général, pour différents critères de sélection de réponse.

Les meilleurs résultats sur Alzheimer sont obtenus en combinant le poids des phrases et le poids des réponses calculé en tenant compte des énoncés. Cela correspond au fait que le poids des phrases permet de classer la phrase pertinente dans les premières, et qu'un poids indépendant de la structure des phrases ne permet pas de sélectionner la bonne réponse. Les mesures tenant compte des relations entre réponse et mots de la question dans l'énoncé représentées par la distance d'édition sur les arbres ou l'alignement permettent en revanche de mieux sélectionner la réponse correcte. Nos meilleurs résultats sur la tâche Alzheimer correspondent à un c@1 de 0,39 et place le système en 3ème position sur 7. Lors de QA4MRE 2012, le meilleur système a obtenu un c@1 de 0,55 et le deuxième 0,47. L'analyse des erreurs sur cette tâche montre que le principal problème réside dans la variabilité lexicale des expressions.

Les résultats obtenus sur la tâche principale ont été calculés afin de montrer la stabilité du système. Notons que cette tâche comporte des questions nécessitant d'utiliser la collection de documents pour sélectionner la réponse correcte. Les meilleures performances sont obtenues en utilisant le poids des réponses sur les termes communs, dans la meilleure phrase. La fréquence donne des résultats moins bons sur ce corpus. Une explication possible de la moins bonne performance des distances d'édition sur la tâche principale est que pour cette dernière, les réponses sont généralement plus longues que sur Alzheimer, ce qui est plus difficile à prendre en compte dans les distances et mieux pris en compte avec le recouvrement lexical. Ces résultats placent le système en 3ème position sur 11 dans cette tâche (1er 0,65 et 2ème 0,40).

4. Conclusion

Dans cet article, nous avons présenté l'adaptation de notre système de questions-réponse QALC pour une tâche de réponse à des questions de compréhension de textes. La sélection des réponses correctes repose sur la reconnaissance d'implication textuelle entre les hypothèses et les textes. Nous avons adapté différentes mesures de reconnaissance de l'implication textuelle reposant à la fois sur les formes de surface et les arbres syntaxiques des hypothèses et phrases réponses.

Les meilleurs résultats sont obtenus en combinant la pondération des phrases sur les termes communs et leur proximité, et des mesures de distance syntaxique pour sélectionner la réponse. Cette combinaison permet de tenir compte du recouvrement lexical qui est un bon critère pour sélectionner des phrases pertinentes, et de la structure de la phrase pour sélectionner la bonne réponse.

Ces résultats pourraient être améliorés avec une meilleure prise en compte de variantes de termes, par exemple en calculant des proximités sémantiques entre termes de l'hypothèse et des phrases, en fonction de leurs contextes d'apparition en corpus.

5. Bibliographie

Baker C., Fillmore C., Lowe J., « The berkeley framenet project », *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, Association for Computational Linguistics, p. 86-90, 1998.

Bhaskar P., Pakray P., Banerjee S., Banerjee S., Bandyopadhyay S., Gelbukh A., « Question Answering System for QA4MRE@CLEF 2012 », *CLEF Workshop on QA4MRE*, 2012.

Clark P., Harrison P., Yao X., « An Entailment-Based Approach to the QA4MRE Challenge », *Proc. CLEF 2012 (Conference and Labs of the Evaluation Forum)*, 2012.

Dagan I., Glickman O., Magnini B., « The pascal recognising textual entailment challenge », *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*p. 177-190, 2006.

de Chalendar G., Dalmas T., Elkateb-Gara F., Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I., Vilnat A., « The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet », *TREC11 NIST special publication SP*, 2002.

Grappy A., Grau B., « Validation du type de la réponse dans un système de questions réponses », *Document numérique*, vol. 14, n° 2, p. 125-147, 2011.

Jacquemin C., « Syntagmatic and paradigmatic representations of term variation », *Proceedings of the 37th annual meeting of ACL*, 1999.

Levy R., Andrew G., « Tregex and Tsurgeon : tools for querying and manipulating tree data structures », *in* , ELDA (ed.), *Fifth international conference on Language Resources and Evaluation (LREC 2006)*, ELDA, Genoa, Italy, 2006.

McClosky D., Any Domain Parsing : Automatic Domain Adaptation for Natural Language Parsing, PhD thesis, Brown University, 2010.

Pakray P., Bhaskar P., Banerjee S., Pal B., Bandyopadhyay S., Gelbukh A., « A Hybrid Question Answering System based on Information Retrieval and Answer Validation », *CLEF 2011 Workshop on QA4MRE*, 2011.

Peñas A., Rodrigo Á., Sama V., Verdejo F., « Overview of the answer validation exercise 2006 », *Evaluation of Multilingual and Multi-modal Information Retrieval*p. 257-264, 2007.

Penas A., Rodrigo Á., Verdejo F., « Overview of the answer validation exercise 2007 », *Advances in Multilingual and Multimodal Information Retrieval*p. 237-248, 2008.

Pho V.-M., « Génération de réponses pour un système de questions-réponses », *RJCRI (CORIA)*, 2012.

Rodrigo Á., Peñas A., Verdejo F., « Overview of the answer validation exercise 2008 », *Evaluating Systems for Multilingual and Multimodal Information Access*p. 296-313, 2009.

Snover M., Madnani N., Dorr B., Schwartz R., « TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate », *Machine Translation*, vol. 23, n° 2, p. 117-127, 2009.

Wang K., Ming Z., Chua T., « A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services », *Proceedings of the 32nd international ACM SIGIR Conference*, ACM, p. 187-194, 2009.

Zhang K., Shasha D., « Simple fast algorithms for the editing distance between trees and related problems », *SIAM J. Comput.*, vol. 18, n° 6, p. 1245-1262, 1989.