Étude de l'influence de la taille du corpus de référence sur l'extraction terminologique automatique contrastive

Audrey Laroche, Patrick Drouin, Gabriel Bernier-Colborne
OLST, Département de linguistique et de traduction, Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal (Québec) Canada H3C 3J7
{audrey.laroche,patrick.drouin,gabriel.bernier-colborne}@umontreal.ca

Résumé

L'extraction terminologique automatique contrastive (ETAC), où les candidats termes sont repérés au moyen de la comparaison de leurs fréquences dans des corpus techniques et non techniques, est influencée par plusieurs paramètres. L'un des plus importants est la taille du corpus non technique, car ce dernier demeure le même d'une extraction à l'autre. À partir de ressources en anglais liées aux domaines de l'automobile et du terrorisme, nous démontrons expérimentalement que les corpus non techniques qui donnent les meilleures extractions sont significativement plus courts que ceux utilisés dans les systèmes d'ETAC jusqu'à maintenant.

1 Introduction

L'extraction terminologique automatique a tour à tour été étudiée, depuis une vingtaine d'années, selon des approches linguistiques, statistiques ou hybrides, ces dernières étant de loin les plus fréquentes (Daille, 1994; Kageura et Umino, 1996; Kit, 2002; Drouin, 2003; Vu et al., 2008; Lassalle, 2011). Les techniques d'extraction sont bien documentées et des extracteurs terminologiques sont disponibles autant pour la recherche que pour le privé. Les chercheurs s'intéressent à présent à la qualité de l'extraction et à l'influence des ressources qui lui sont nécessaires.

L'une des approches les plus répandues, que nous appelons Extraction Terminologique Automatique Contrastive (ETAC), repose sur la comparaison des fréquences des candidats termes (CT) dans deux corpus de natures différentes (l'un est général, l'autre spécialisé) dans le but de déterminer la probabilité que les CT soient des termes valides. Les CT sont ainsi caractérisés selon leur potentiel terminologique (*termhood)*, c'est-à-dire le degré de spécialisation de leur sens dans le domaine à l'étude (Kageura et Umino, 1996). Pour mesurer ce potentiel, Damereau (1993) utilise le ratio des fréquences relatives de deux corpus ; Ahmad et al. (1994) et Chung (2003) utilisent des calculs basés sur les fréquences normalisées, tandis que l'approche de (Drouin, 2003) est basée sur les probabilités. Drouin et Langlais (2006) comparent expérimentalement la performance de diverses mesures du potentiel terminologique : fréquence brute, spécificité (Lafon, 1980), χ^2, ratio de vraisemblance (Dunning, 1993) et modèle de langue trigramme. Les meilleures extractions sont obtenues avec la fréquence brute, suivie de près par la spécificité. Cette dernière permet de recenser des CT plus longs, ce qui lui confère un avantage étant donné que les termes ont tendance à être complexes dans les domaines techniques (Drouin et Langlais, 2006).

Ces techniques d'ETAC dépendent fortement des ressources utilisées. Dans ce travail, des expériences sont menées en utilisant des corpus de tailles variées de façon à étudier l'influence de la taille sur les résultats de l'extraction terminologique. Il importe de trouver la taille optimale du corpus non technique parce que les fréquences de celui-ci sont précompilées (seul le corpus spécialisé varie d'une extraction à l'autre).

Dans la section 2, nous décrivons l'approche d'ETAC. Les ressources et le protocole expérimentaux figurent aux sections 3 et 4 ; les résultats suivent à la section 5. À la section 6, nous faisons le lien entre nos découvertes et les travaux antérieurs et présentons les travaux à venir.

Long papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 66–72
Paris, 8–10 November 2011

2 Approche

L'une des stratégies d'extraction de termes simples et complexes possibles est l'ETAC, qui est basée sur la comparaison de la fréquence des mots dans un corpus non technique (le *corpus de référence*, CR) et un corpus technique lié à un domaine particulier (le *corpus d'analyse*, CA). L'ETAC repose sur l'hypothèse selon laquelle les termes spécifiques au domaine ont des fréquences significativement plus élevées dans le CA que les fréquences attendues selon le CR. Cette technique utilise une approche hybride, basée sur des indices linguistiques et statistiques (Drouin, 2003).

Pour extraire des termes, la distribution des unités lexicales (noms, adjectifs, verbes, adverbes et syntagmes nominaux) dans le corpus de référence est d'abord modélisée au moyen de leur fréquence [1]. Les syntagmes sont identifiés à l'aide d'une grammaire simple composée d'expressions régulières formant des patrons de parties du discours [2].

Ensuite, les candidats termes (CT) simples et complexes sont extraits du CA à l'aide de la grammaire [3], qui s'inspire de la description que fait (Sager, 1990) des structures syntaxiques des termes.

	CR	**CA**	**Total**
Fréquence du CT	a	b	$a+b$
Fréquence des autres unités lex.	c	d	$c+d$
Total	$a+c$	$b+d$	$N=a+b+c+d$

TABLE 1: Table de contingence

Enfin, le potentiel terminologique de chaque CT est calculé à l'aide d'une mesure statistique, la spécificité (Lafon, 1980), qui quantifie la déviation de la fréquence du CT dans le CA par rapport à la valeur attendue selon les fréquences calculées dans le CR. Le calcul est basé sur les valeurs d'une table de contingence (Table 1) (Drouin et Langlais, 2006) :

$$\begin{aligned} log\ P(X{=}b) = & log(a+b)! + log(c+d)! \\ & + log(b+d)! + log(a+c)! \\ & - log(N)! - log(b)! - log(a)! \\ & - log(b)! - log(c)! \end{aligned}$$

Dans le calcul de la spécificité, la variable aléatoire X suit une loi hypergéométrique qui, selon (Lafon, 1980), est mieux adaptée que les distributions normale ou de Poisson à la population discrète que forment les unités lexicales et à leurs fréquences dispersées. Plus un candidat terme a une spécificité élevée, plus il est susceptible d'être un terme valide. Un seuil minimal de 3,09 est fixé : si la spécificité d'un CT est inférieure à ce seuil, alors il n'est pas inclus dans la liste des candidats termes. La valeur 3,09 signifie que la probabilité d'observer la fréquence du CT dans le CA est moins de 1/1000 par rapport à sa fréquence dans le CR (Lebart et Salem, 1994, p. 83). Étant donné qu'il s'agit d'une proportion, ce seuil demeure en principe adéquat pour toute taille de CR.

L'algorithme d'extraction de termes décrit ci-dessus est implémenté dans le logiciel TermoStat [4] (Drouin, 2003), qui prend en entrée un CA et produit une liste de CT ordonnés selon leur score de potentiel terminologique. Dans l'implémentation originale de TermoStat, le CR est constitué d'articles du quotidien montréalais *The Gazette* totalisant 7,4 millions d'occurrences. Selon (Drouin, 2003), cette taille « modeste » devrait être « maximisée » pour assurer la stabilité des CT identifiés par l'algorithme. Ceci rejoint l'hypothèse plus vaste de (Church et Mercer, 1993) selon laquelle « *more data are better data* ». La taille du CR est un facteur important pour la qualité de l'extraction terminologique, puisqu'elle a une influence directe sur le modèle de la distribution des unités lexicales de la langue générale qui est comparé au CA. Dans les extracteurs comme TermoStat, ce modèle étant précompilé avant que les utilisateurs ne soumettent leur propre CA, il est pertinent de déterminer la taille optimale du CR.

1. Plus précisément, ce sont les fréquences des lemmes qui sont calculées.

2. La grammaire est consultable à : `http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html#fonctionnement`.

3. À cette étape, si, dans le CA, un candidat terme (CT_1) est inclus dans un candidat plus long (CT_2) et n'apparaît jamais en-dehors du CT_2, alors le CT_1 est considéré comme un fragment de terme et est exclu de la liste des CT. Dans la présente étude, seuls les termes nominaux sont extraits en raison des terminologies de référence disponibles (section 3).

4. `http://olst.ling.umontreal.ca/~drouinp/termostat_web/`. TermoStat peut extraire des CT de corpus anglais, espagnols, français, italiens et portugais.

3 Ressources

L'algorithme d'ETAC est basé sur deux types de corpus : un corpus de référence (CR), qui représente la langue générale, et un corpus d'analyse de langue spécialisée (CA) duquel sont extraits les candidats termes (CT). Les corpus sont lemmatisés et leurs parties du discours sont étiquetées avec TreeTagger [5] avant leur soumission à TermoStat.

3.1 Corpus de référence

Le corpus de référence est le *British National Corpus* (BNC) [6], qui est constitué de documents de natures variées en anglais britannique contemporain totalisant 100 millions d'occurrences. Le BNC est un CR approprié pour l'ETAC puisqu'il représente la langue générale : l'hétérogénéité des domaines, des genres et des auteurs compense les segments qui pourraient être techniques.

3.2 Corpus d'analyse et terminologies de référence

Les corpus d'analyse proviennent des domaines de l'automobile (CA_1) et du terrorisme (CA_2). Les extractions terminologiques à partir du CA_1 et du CA_2 sont évaluées selon leur terminologie de référence respective qui, dans les deux cas, a été constituée par des terminologues à partir des corpus même [7].

Les terminologies de référence ont été construites dans un but terminologique et ne sont pas spécifiquement conçues pour évaluer l'extraction. Les critères de sélection des termes sont très précis. Dans le cas du CA_1, plusieurs termes ont été laissés de côté parce qu'ils sortaient des frontières du domaine tel que défini par les terminologues ; ces termes seraient normalement considérés comme valides dans une évaluation manuelle de l'extraction. Dans le cas du CA_2, seuls les termes faisant partie de certaines classes conceptuelles d'un sous-domaine du terrorisme ont été retenus. Ces décisions quelque peu atypiques de sélection des termes de référence ont un impact négatif sur la précision de l'extracteur terminologique, celui-ci ne pouvant choisir les mêmes termes qu'un terminologue humain.

Domaine de l'automobile : Le corpus du domaine de l'automobile (CA_1) est constitué de trois manuels de mécanique automobile et comprend 225 000 occurrences.

Le corpus a été analysé manuellement par des terminologues qui ont annoté chaque occurrence des termes (noms ou syntagmes nominaux) en fonction de critères de sélection spécifiques, notamment les critères lexicosémantiques de (L'Homme, 2004, p. 64–66). Le critère fondamental est le potentiel terminologique. Les candidats qui sont morphologiquement ou paradigmatiquement liés à un terme valide sont eux-mêmes valides.

En plus des termes de base, les terminologues ont annoté les synonymes et les variations de termes, c'est-à-dire les variantes graphiques (*sparkplug*/*spark plug*), les acronymes et les réductions anaphoriques (quand un terme complexe apparaissant plus tôt dans le texte est remplacé par un de ses composants, habituellement la tête). Des instructions particulières régissaient l'annotation des termes coordonnés (*intake and exhaust valves*) ou séparés par une ponctuation ou des syntagmes enchâssés (*EGR (exhaust gas recirculation) valve*).

La terminologie de référence pour le domaine de l'automobile contient 1571 termes simples et 3918 termes complexes. Parmi ces 5489 termes, 2668 sont des termes de base, tandis que 2821 sont des variantes terminologiques ou des synonymes.

Domaine du terrorisme : Le second domaine que nous utilisons dans nos expériences est lié au projet SACOT [8], portant sur la construction semi-automatique d'ontologies à partir de textes. Le corpus de 50 000 occurrences est formé de documents portant sur le domaine du terrorisme, et plus particulièrement sur celui des engins explosifs improvisés (EEI).

Tout comme dans le cas des ressources du domaine automobile, les critères de sélection de termes de (L'Homme, 2004) ont été suivis. Seuls les noms et les syntagmes nominaux référant aux objets du domaine des EEI sont sélectionnés. L'application de ces critères a permis d'obtenir une terminologie de référence comptant 187 termes simples et 610 termes complexes, pour un total de 797 unités.

5. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

6. `http://www.natcorp.ox.ac.uk/`

7. Ces corpus provenant de documents protégés par des droits d'auteur, nous ne pouvons pas rendre publiques les terminologies de référence.

8. SACOT (2006–2009) a été financé par R & D pour la défense Canada (chercheur principal : Alain Auger).

4 Protocole expérimental

Dans chaque expérience, l'extracteur terminologique analyse un corpus spécialisé en utilisant un corpus de référence d'une taille donnée. Les candidats termes extraits sont ensuite comparés à la terminologie de référence selon des mesures d'évaluation classiques et terminologiques afin de déterminer la taille optimale du CR.

4.1 Tailles des corpus de référence

Le corpus de référence initialement utilisé dans TermoStat compte 8 millions d'occurrences. Nous évaluons la performance de l'extraction à partir de corpus de 0,1, 0,2, 0,5, 0,8, 1, 2, 4, 6, 8, 10, 12, 14, 16 et 32 millions d'occurrences. Chacune de ces tailles est testée trois fois, en utilisant trois partitions distinctes du BNC, de sorte qu'aucun segment ne soit identique entre les essais. Cette procédure permet de diminuer les effets de la variation du contenu des CR pour se concentrer uniquement sur l'influence de la taille. Les résultats présentés à la section 5 sont en fait les moyennes des mesures d'évaluation pour les trois essais.

4.2 Mesures d'évaluation

La performance des extractions est évaluée au moyen de la précision (*P*), du rappel (*R*) et de la F-mesure (*F*). La *P* est le nombre de CT présents dans la terminologie de référence divisé par le nombre de CT ; le *R* est le nombre de CT présents dans la terminologie de référence divisé par le nombre de termes de référence. *F* est la moyenne harmonique de *P* et *R* ($F = \frac{2\times(P\times R)}{(P+R)}$).

La performance est aussi évaluée avec les mesures terminologiques proposées par (Nazarenko et al., 2009) à l'aide de l'outil Termometer [9]. Ces mesures tiennent compte de la *pertinence graduelle* des termes : sont considérés partiellement valides les CT qui, bien que n'apparaissant pas dans la terminologie de référence, sont similaires à l'un des termes de référence (par ex. *base de données relationnelle* et *base de données*). Cette pertinence graduelle s'exprime au moyen de la *distance terminologique*, qui combine les distances de Levenshtein normalisées aux niveaux 1) des caractères dans les chaînes et 2) des mots dans les termes complexes. Si la distance entre un CT et un des termes de référence est inférieure à un seuil τ, alors ce CT est pertinent [10]. Pour tenir compte de la *variabilité* (granulaire) *de la terminologie de référence*, la correspondance optimale *P(O)* entre la liste des CT et la terminologie de référence est calculée. La précision (*PT*) et le rappel (*RT*) terminologiques sont définis ainsi :

$$PT = \frac{Pert(O,G)}{|P(O)|} \qquad RT = \frac{Pert(O,G)}{|G|}$$

où $Pert(O,G)$ représente la pertinence globale de la liste de CT par rapport à la terminologie de référence, $|P(O)|$ est le nombre de termes dans la correspondance optimale et $|G|$ est la taille de la terminologie de référence. La F-mesure terminologique (*FT*) est la moyenne harmonique de PT et RT.

Dans le corpus du domaine automobile, toutes les variantes morphologiques (singulier/pluriel) sont annotées et liées à leur forme canonique dans la terminologie de référence. Le corpus du domaine du terrorisme n'étant pas annoté, seules les formes canoniques se retrouvent dans la terminologie de référence. Puisque l'extracteur TermoStat liste les formes canoniques et leurs variantes, deux stratégies d'évaluation sont utilisées. Pour le CA_1, les métriques s'appliquent à toutes les variantes trouvées par TermoStat. Pour le CA_2, seules les formes canoniques identifiées par TermoStat sont prises en compte.

5 Résultats

5.1 Quantité de candidats termes

La première étape de l'ETAC est la modélisation de la distribution des mots et des syntagmes nominaux (SN) dans la langue générale en comptant les fréquences dans un CR (section 2). La Figure 1 montre le nombre de mots simples (partie inférieure des bandes) et de syntagmes nominaux (partie supérieure) distincts calculé pour chaque taille de CR. La quantité de SN croît plus rapidement que le nombre de mots lorsque la taille du CR croît, ce qui est normal du point de vue combinatoire.

La Figure 2 montre le nombre de CT simples et complexes extraits des deux CA pour chaque taille de CR. Le nombre de CT dans le CA_1 (celui de l'automobile) se stabilise lorsque le CR at-

9. `http://sourceforge.net/projects/termometerxd/`

10. Dans nos expériences, le seuil optimal est calculé à l'aide de Termometer : $\tau = 0.536618$ pour la terminologie de l'automobile et $\tau = 0.618029$ pour celle du terrorisme.

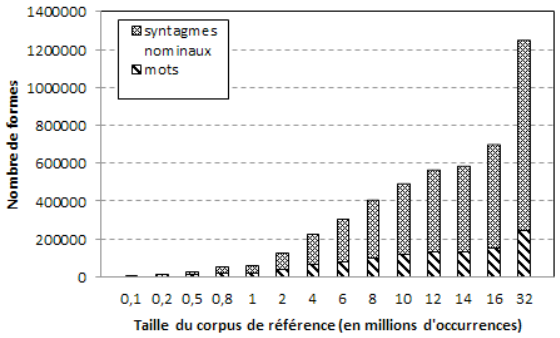

FIGURE 1: Nombre d'unités lexicales extraites du CR

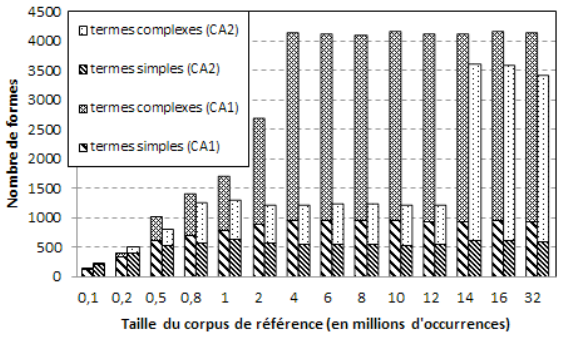

FIGURE 2: Nombre de termes extraits des CA

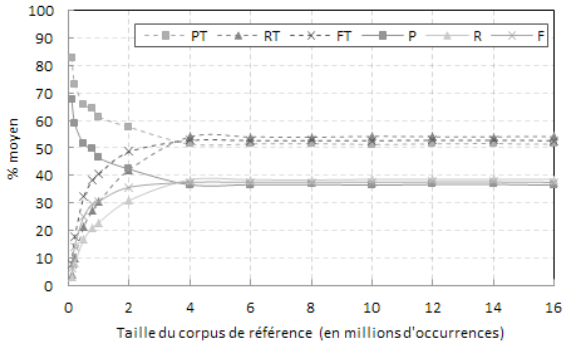

FIGURE 3: Performances des extractions (CA_1)

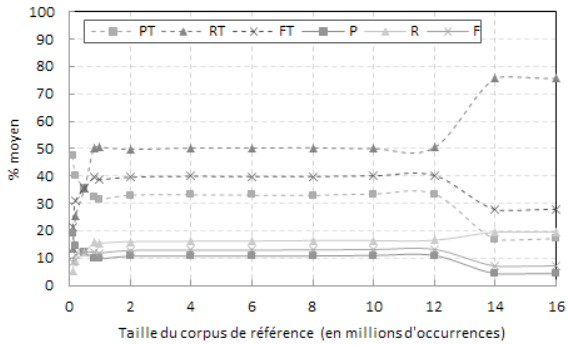

FIGURE 4: Performance des extractions (CA_2)

teint 4 millions d'occurrences, et ce, même si les CR de tailles supérieures contiennent plus d'unités linguistiques distinctes. La quantité de CT extraits du CA_2 atteint un premier plateau avec le CR de 800 000 occurrences, puis un second à partir de 14 millions d'occurrences (plus de 2000 nouveaux CT sont trouvés). À ce point, le nombre de termes contenant entre 2 et 5 composants augmente fortement et des patrons syntaxiques plus longs sont identifiés. Environ 160 CT complexes sont perdus lorsque le CR passe de 16 à 32 millions d'occurrences ; seulement 7 d'entre eux sont des termes de référence, mais davantage seraient considérés valides si le domaine (les EEI) était moins restreint, comme *blast area*. Il serait intéressant de vérifier dans de futures expériences s'il existe une taille maximale de CR au-delà de laquelle l'extraction est significativement perturbée.

5.2 Qualité de l'extraction

Les Figures 3 et 4 montrent la qualité de l'extraction des CT dans les CA_1 et CA_2 en fonction de la taille du CR selon les mesures classiques et terminologiques de précision, rappel et F-mesure (P, R, F, PT, RT, FT). Dans les deux figures, les mesures classiques et terminologiques suivent la même courbe, mais les scores terminologiques sont plus élevés ; ce comportement est conforme aux fondements théoriques des mesures terminologiques de (Nazarenko et al., 2009).

Pour le CA_1, les performances atteignent un plateau avec le CR de 4 millions d'occurrences (FT de 52,8 %), après une rapide convergence où la P diminue et le R augmente. Les CR plus longs n'ont pas d'effet sur la qualité de l'extraction. Pour le CA_2, un premier plateau est atteint avec le CR de 800 000 occurrences, et un second avec celui de 14 millions d'occurrences [11]. Les performances du CA_2 (FT maximale de 39 % atteinte avec le CR de 800 000 occurrences) sont significativement inférieures à celles du CA_1. Ceci s'explique par le fait que le rappel maximal classique, pour le CA_2, est de 66,2 %, car les variantes morphologiques ne sont pas annotées dans le corpus et la terminologie de référence ne contient que des formes canoniques. Le rappel est aussi influencé par le fait que la nature du domaine fait en sorte que plusieurs de ses termes appartiennent aussi à la langue générale (les EEI peuvent être construits à partir d'objets

11. Ce plateau continue jusqu'au CR de 32 millions d'occurrences (absent du graphique faute d'espace).

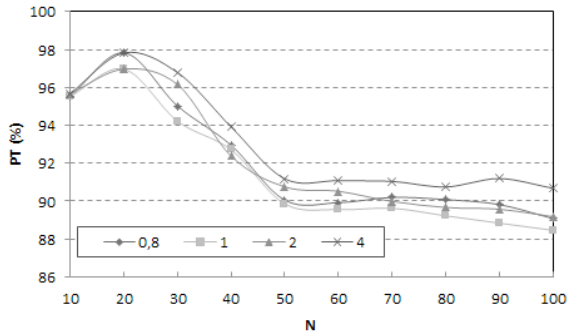

FIGURE 5: PT des N premiers CT (CA_1)

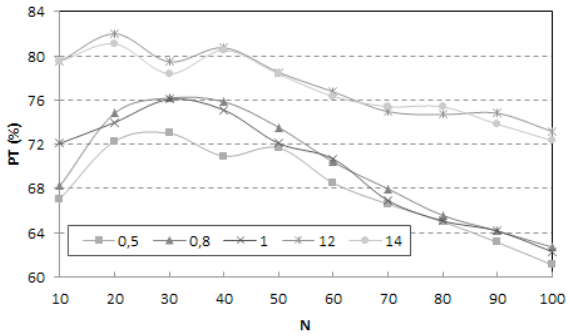

FIGURE 6: PT des N premiers CT (CA_2)

variés comme des bouteilles, des portes, etc.).

Les plateaux apparaissant dans les Figures 3 et 4 correspondent aux limites de nombre de CT extraits pour les deux CA (Figure 2). Ils correspondent de plus (à l'exception du second plateau du CA_2) aux cas où la taille du CA est d'environ 6 % celle du corpus de référence [12]. L'existence de ces plateaux indique que le CR n'a pas à être arbitrairement long, contredisant la proposition de (Drouin, 2003) de le maximiser.

5.3 Qualité des N premiers candidats termes

Les terminologues s'intéressent habituellement aux CT ayant les scores de potentiel terminologique les plus élevés et non à la liste de tous les CT extraits (Table 2 ; les termes de référence sont en italiques). Les Figures 5 et 6 montrent les précisions terminologiques en considérant les 10 à 100 plus forts CT pour certaines tailles de CR [13].

CA	CR	10 premiers CT
1	1 M	*engine*, *system*, *valve*, *fuel*, *car*, pressure, *oil*, *air*, *brake*, *vehicle*
	4 M	*engine*, *valve*, *fuel*, *car*, *system*, *brake*, *oil*, pressure, *cylinder*, *piston*
2	0,8 M	device, ieds, *vehicle*, *circuit*, enemy, *attack*, force, bomb, threat, explosives
	14 M	ieds, device, *convoy*, explosives, *circuit*, *explosive device*, *vehicle*, *detonation*, *armor*, theater

TABLE 2: Exemples de N premiers CT (N=10)

Pour le CA_1, plus de 88 % des 100 plus forts CT sont des termes de référence, peu importe la taille du CR ; la plus haute PT (90,7 % avec N=100) est atteinte avec le CR de 4 millions d'occurrences.

Les PT atteintes avec le CA_2 sont plus faibles. Les CR les plus petits produisent des PT d'environ 62 % (avec N=100), le meilleur étant celui de 0,8 millions d'occurrences ; les plus gros (12 et 14 millions d'occurrences) sont significativement meilleurs (≈73 %, N=100). Malgré tout, la chute de précision avec les CR plus gros lorsque la liste complète de CT est considérée (section 5.2, Figure 4) est trop importante pour qu'ils puissent être désignés comme les meilleurs.

Notre protocole d'évaluation étant entièrement automatique, il est difficile de comparer nos résultats à ceux obtenus par exemple par (Nazarenko et al., 2009) et (Drouin, 2003), où des terminologues examinent les CT. Le protocole utilisé pour le CA_1 ressemble davantage à une évaluation humaine parce qu'il tient compte des variantes morphologiques, contrairement aux évaluations menées avec le CA_2. Les résultats obtenus avec le CA_1 sont donc plus représentatifs de la performance réelle de TermoStat.

6 Conclusion

Nos résultats confirment que la taille du corpus de référence influence la quantité de candidats termes extraits ainsi que leur qualité. Ce paramètre commun aux logiciels basés sur l'ETAC n'avait jamais été étudié malgré son importance. De façon surprenante, contrairement à l'hypothèse de (Drouin, 2003) et à la recommandation de (Church et Mercer, 1993) (dont la tâche consiste à identifier des cooccurrents), un corpus plus gros ne donne pas nécessairement de meilleurs résultats. Parmi les 14 tailles testées, les CR qui permettent d'atteindre les performances maximales comptent

12. Cette observation invite à de plus amples expériences pour déterminer l'influence de la taille du corpus d'analyse, que nous n'avons pas contrôlée dans la présente étude.

13. Ces tailles sont les points critiques où les performances varient significativement dans les Figures 3 et 4.

respectivement 4 (dans le cas du CA du domaine de l'automobile) et 0,8 (dans le cas du CA sur le terrorisme) millions d'occurrences, soit beaucoup moins que le CR original de 8 millions d'occurrences utilisé dans TermoStat.

Au-delà d'une certaine taille, il semble y avoir un surapprentissage qui fait en sorte que, même si davantage de phénomènes de la langue générale sont modélisés (Church et Mercer, 1993), l'extraction de CT n'est plus influencée. Un examen plus approfondi des CT extraits selon la taille du CR fournirait une explication plus complète.

Nos résultats sont à rapprocher de ceux de (Zesch et Gurevych, 2010), qui ont découvert, en utilisant 13 versions successives de Wikipédia, que la croissance de ce corpus n'a pas d'effets significatifs sur la performance dans une tâche de similarité sémantique. Selon eux, des corpus plus petits (donc moins coûteux) peuvent être utilisés sans nuire à la performance.

Pour améliorer les logiciels d'ETAC, un CR pourrait être sélectionné parmi plusieurs CR précompilés de façon à ce que le CA fasse environ 6 % de sa taille. Une autre méthode (plus coûteuse) consisterait à mener des extractions en utilisant des CR de plus en plus grands jusqu'à ce que le nombre de CT extraits se stabilise.

Nos prochaines expériences porteront sur l'influence du contenu du corpus de référence sur la qualité de l'extraction et sur la relation entre le choix du seuil de spécificité et la taille du corpus. Les types d'erreurs produites par l'extracteur en fonction de la taille du corpus seront aussi analysés selon des critères terminologiques.

Remerciements

Nous remercions Alain Auger qui nous a permis d'utiliser les données du projet SACOT, Thibault Mondary pour l'outil Termometer, Jonathan Mérel pour son aide avec QT et les relecteurs anonymes pour la pertinence de leurs commentaires.

Références

Ahmad, Khurshid, Andrea Davies, Heather Fulford et Margaret Rogers. 1994. What's in a term ? The semi-automatic extraction of terms from text. *Translation Studies. An Interdiscipline*, Amsterdam/Philadelphia : John Benjamins, p. 267–278.

Chung, Teresa Mihwa. 2003. A Corpus Comparison Approach for Terminology Extraction. *Terminology*, 9(2), p. 221–246.

Church, Kenneth W. et Robert L. Mercer. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, 19(1), p. 1–24.

Damereau, Fred J. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management : an International Journal*, 9(2), p. 433–447.

Daille, Béatrice. 1994. Study And Implementation Of Combined Techniques For Automatic Extraction Of Terminology. *Workshop On The Balancing Act : Combining Symbolic And Statistical Approaches To Language*, p. 29–36.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), p. 99–115.

Drouin, Patrick et Philippe Langlais. 2006. Évaluation du potentiel terminologique de candidats termes. *8es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2006)*, p. 379–388.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), p. 61–74.

Kageura, Kyo et Bin Umino. 1996. Methods for automatic term recognition : A review. *Terminology*, 3(2), p. 259–289.

Kit, Chunyu. 2002. Corpus Tools for Retrieving and Deriving Termhood Evidence. *Proceedings of the 5th East Asia Forum of Terminology*, p. 69–80.

L'Homme, Marie-Claude. 2004. *La terminologie : principes et techniques*, Montréal : Presses de l'Université de Montréal.

Lafon, Pierre. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, 1(1), p. 128–165.

Lassalle, Edmond. 2011. Acquisition automatique de terminologie à partir de corpus de texte. *Traitement Automatique des Langues Naturelles (TALN 2011)*.

Lebart, Ludovic et André Salem. 1994. *Statistique textuelle*, Paris : Dunod.

Nazarenko, Adeline, Haïfa Zargayouna, Olivier Hamon et Jonathan van Puymbrouck. 2009. Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement automatique des Langues*, 50(1), p. 257–281.

Sager, Juan C. 1990. *A Practical Course in Terminology Processing*, Amsterdam/Philadelphia : John Benjamins.

Vu, Thuy, Ai Ti Aw et Min Zhang. 2008. Term Extraction Through Unithood and Termhood Unification. *Third International Joint Conference on Natural Language Processing (IJNLP 2008)*, p. 631–636.

Zesch, Torsten et Iryna Gurevych. 2010. The More the Better ? Assessing the Influence of Wikipedia's Growth on Semantic Relatedness Measures. *Language Resources and Evaluation (LREC 2010)*, p. 1374–1380.