

Extraction automatique: éléments pour la construction d'un SEACUSE (Système d'Extraction Automatique de Candidats à Unités de Sens Spécialisé), Rosa Estopà, Thèse de doctorat en linguistique appliquée, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, juillet 1999. Directeur de thèse: M.Teresa Cabré. Jury: Toni Badia (président), Christian Jacquemin (secrétaire), Horacio Rodríguez, Josep Lluís Borona, Mercè Lorente.

Notre thèse présente une étude linguistique sur les unités de sens spécialisé dans le domaine médical visant à leur reconnaissance et extraction automatique à partir de textes spécialisés. C'est une étude qui a permis de proposer une maquette de système d'extraction automatique de candidats à unités de sens spécialisé (SEACUSE) à des fins d'applications professionnelles.

La thèse s'inscrit dans le cadre des recherches sur le lexique qui se déroulent à l'Institut Universitari de Lingüística Aplicada (IULA) de l'Universitat Pompeu Fabra; concrètement elle fait partie d'un projet de reconnaissance et extraction automatique de terminologie dont l'un des objectifs est l'élaboration d'un extracteur pour le catalan. C'est un travail qui a été élaboré sous la perspective linguistique de la terminologie et qui part concrètement de son approche communicative.

Les objectifs généraux de la thèse sont les trois suivants:

1. Analyser et évaluer le fonctionnement des outils d'extraction automatique de candidats à terme actuels dans le but d'avoir des éléments pour dessiner un extracteur pour le catalan.
2. Proposer des éléments de reconnaissance des unités de sens spécialisé (USE) pertinentes pour les extraire automatiquement des textes.
3. Dessiner une maquette de Système d'Extraction Automatique de Candidats à Unités de Sens Spécialisé (SEACUSE) adéquat aux besoins professionnels des différents collectifs d'utilisateurs.

Le travail s'articule en trois parties: la première présente une analyse et une critique évaluative du fonctionnement des extracteurs existants par rapport au dépouillement manuel –chapitres 1, 2, 3, et 4–; la deuxième décrit l'objet d'extraction et propose un extracteur multifonctionnel –chapitres 5, 6 et 7–; finalement, la troisième partie inclut les conclusions et la bibliographie de référence –chapitres 8 et 9.

Le premier chapitre décrit et analyse les principaux outils d'extraction automatique de candidats à terme (SEACAT) dans le but d'exposer l'état de la question et de mettre en évidence les qualités et notamment les faiblesses de chacun des outils. L'étude constate que, si les outils fonctionnent correctement, ils ne sont pas absolument satisfaisants et qu'on peut donc les améliorer.

Dans le deuxième chapitre, nous montrons d'abord, sur la base de corpus textuels spécialisés, la validité des patrons structuraux des unités terminologiques polylexicales que nous avons établis à l'aide de corpus lexicographiques dans un travail de doctorat antérieur. Ensuite, nous mettons en évidence, à partir de plusieurs dépouillements d'un même corpus textuel, les principales limitations des SEACAT qui se basent sur des patrons morpho-syntaxiques. Ces limitations se manifestent sous deux aspects: le silence (unités pertinentes non détectées par l'extracteur) et le bruit (unités non pertinentes présentées comme si elles l'étaient).

Nous étudions le silence et le bruit de manière systématique dans le troisième et le quatrième chapitres, respectivement. Ainsi, d'abord nous analysons les types et les causes du silence que produisent les SEACAT, et ensuite les types et les causes du bruit généré par ces systèmes.

Le cinquième chapitre propose des éléments et des stratégies pour qu'un outil d'extraction automatique puisse réduire le silence et le bruit, et ainsi parvenir à des résultats qui se rapprochent davantage de la reconnaissance et de la délimitation manuelle des unités de sens spécialisé. Dans ce chapitre, après avoir valorisé les dépouillements que des spécialistes ont faits d'un corpus textuel de médecine, nous proposons d'élargir l'objet d'extraction. Étant donné que les textes spécialisés présentent beaucoup d'unités qui ne sont pas des unités terminologiques, mais qui sont aussi intéressantes du point de vue spécialisé, nous sommes partis d'une unité de sens spécialisé large. Les USE véhiculent de la connaissance spécialisée et, du point de vue de la forme, elles comprennent plusieurs types d'unités signiques, tant linguistiques que non linguistiques. Les unités linguistiques peuvent être aussi bien lexicales que syntaxiques; elles peuvent également appartenir à différentes catégories grammaticales.

Le sixième chapitre introduit le point de vue fonctionnel d'une unité: nous partons du fait que la pertinence d'une unité de sens spécialisé dépend des besoins professionnels générés par une activité déterminée. Nous avons ainsi remarqué que toutes les activités professionnelles n'ont pas besoin du même type ni du même nombre

d'unités spécialisées. Cette hypothèse est vérifiée par une preuve expérimentale basée sur les besoins de quatre activités professionnelles différentes.

Finalement, le septième chapitre présente une proposition de maquette de SEACUSE qui, en plus des stratégies exposées dans le cinquième chapitre, tient compte des finalités des professionnels lors de la présentation des résultats.

La principale contribution appliquée de notre thèse est l'élaboration d'une maquette de SEACUSE, qui améliore les SEACAT fondamentalement dans cinq aspects:

1. Le SEACUSE augmente le nombre d'unités terminologiques détectées, car il reconnaît non seulement les unités polylexicales, mais encore les monolexicales.
2. Il élargit le concept d'objet d'extraction, car au lieu de se fixer seulement sur les unités terminologiques, comme le faisaient les SEACAT, il détecte toutes les unités du texte qui ont un sens spécialisé.
3. Il rend plus précise la reconnaissance des USE, grâce à la combinaison de stratégies de nature différente adaptées à chaque type d'USE.
4. Il rend le dépouillement plus adéquat aux besoins d'une activité professionnelle.
5. Il intègre la reconnaissance et l'extraction d'USE dans d'autres systèmes de traitement du langage plus complexes.

Mais, outre la proposition d'une maquette de SEACUSE, il faut souligner, comme une contribution du travail, la distinction entre Unité Terminologique et Unité de Sens Spécialisé, qui permet de considérer comme un objet d'étude d'autres unités qui prennent une valeur spécialisée; des unités qui sont catégoriellement, syntaxiquement et sémantiquement différentes des termes.

Dans ce sens, nous considérons importante la description linguistique globale que nous avons faite des USE dans des textes spécialisés du domaine spécifique médical, qui contribue à leur distinction et à leur classification. En partant de l'idée que l'analyse linguistique de l'objet d'extraction doit être à la base de la construction d'un extracteur, nous n'avons pas fait une description des USE dans l'abstrait. Au contraire, nous avons souligné que cette description doit servir à reconnaître et à extraire automatiquement les USE des textes spécialisés, de manière plus précise et exhaustive que les systèmes existants. C'est dans ce sens que l'on peut dire que la thèse se situe à la frontière entre la description linguistique et le dessin d'applications informatiques dans le domaine de la terminologie.

Un autre intérêt de notre travail est de considérer la condition de pertinence d'une USE en fonction d'une activité professionnelle déterminée. C'est un concept qui a

permis d'établir des profils professionnels d'USE et qui a mis en évidence le fait que toute application terminologique, pour être utile, ne peut pas fonctionner de la même manière pour des finalités différentes, mais qu'elle doit être adéquate aux besoins professionnels de leurs utilisateurs.

rosa.estopa@trad.upf.es