

D'une hiérarchie figée de concepts vers une hiérarchie évolutive de notions de base

Alain Joubert, Mathieu Lafourcade
{joubert, lafourcade}@lirmm.fr

LIRMM – UM2
Laboratoire d'Informatique, de Robotique
et de Microélectronique de Montpellier
161, rue Ada – 34392 Montpellier Cédex 5 – France

Abstract

In the construction of a thematic representation system of knowledge in Natural Language Processing, it became current to use a thesaurus. Essentially, a thesaurus consists of a set of concepts, often organized in an arborescent structure : it is a fundamental, but relatively fixed tool. But concepts (and language!) are evolving, and this, more and more quickly, particularly in technical fields. We propose a system which makes possible for the notion of concept to evolve by the introduction of the "Basic Notions". Those, necessarily definite on the vector space of the concepts of the thesaurus, constitute another generating system of the space of thematic representation of knowledge. Contrary to the concepts of the thesaurus, the basic notions evolve progressively with the analysis of new texts. We discuss the optimal value of the dimension of the space of representation generated by the basic notions, then of the determination of the acceptations allowing to express them. Lastly, we consider the differentiation between basic notions of general space and those of a specialized field.

Key words : Natural Language Processing, conceptual vectors, basic notions, thesaurus, thematic distance

Résumé

Dans la construction d'un système de représentation thématique des connaissances en Traitement Automatique du Langage Naturel, il est devenu courant d'utiliser un thésaurus. Par essence, un thésaurus est constitué d'un ensemble de concepts, souvent organisé en une structuration arborescente : c'est un instrument fondamental, mais relativement figé. Or les notions (et la langue !) évoluent, et ce, de plus en plus rapidement, en particulier dans les domaines techniques. Nous proposons un système qui permet de faire évoluer la notion de concept par l'introduction des « notions de base ». Celles-ci, définies nécessairement sur l'espace vectoriel des concepts du thésaurus, constituent un autre système générateur de l'espace de représentation thématique des connaissances. Contrairement aux concepts du thésaurus, les notions de base évoluent au fur et à mesure de l'analyse de nouveaux textes. Nous discutons de la valeur optimale de la dimension de l'espace de représentation généré par les notions de base, puis de la détermination des acceptations permettant de les exprimer. Enfin, nous envisageons la différenciation entre notions de base de l'espace généraliste et celles d'un domaine spécialisé.

Mots clés : Traitement Automatique du Langage Naturel, vecteurs conceptuels, notions de base, thésaurus, distance thématique