ANUARI DE FILOLOGIA. ESTUDIS DE LINGÜÍSTICA (Anu.Filol.Est.Lingüíst.)
4/2014, pp. 211-212, ISSN: 2014-1408

PARAPHRASE SCOPE AND TYPOLOGY. A DATA-DRIVEN APPROACH FROM COMPUTATIONAL LINGUISTICS

MARTA VILA RIGAT
marta.vila@ub.edu

S'entén per paràfrasi la igualtat aproximada de significat entre fragments de text que difereixen en la forma. La paràfrasi és omnipresent en les llengües naturals, on es troba expressada de múltiples maneres. D'una banda, la ubiqüitat de la paràfrasi l'ha convertit en el centre d'interès de moltes tasques específiques dins de la lingüística computacional; de l'altra, la seva complexitat ha fet de la paràfrasi un problema que encara no té una solució definitiva.

Dues qüestions bàsiques, lligades a la naturalesa complexa de la paràfrasi, en fan el tractament computacional particularment difícil: l'absència d'una definició precisa i comunament acceptada i la manca de corpus de paràfrasis de referència. Assumint que el coneixement lingüístic ha de ser a la base de la recerca en lingüística computacional, aquesta tesi pretén avançar en dues línies de treball: en la delimitació i comprensió del que s'entén per paràfrasi, i en la creació i anotació de corpus de paràfrasis que proporcionin dades sobre les quals fonamentar tant la recerca com futurs recursos i aplicacions. Amb l'objectiu d'avaluar empíricament el seu potencial, el coneixement i els recursos creats com a resultat d'aquest treball han estat aplicats a la detecció automàtica de plagi.

Aquesta tesi consisteix en un compendi de publicacions i comprèn sis articles principals dividits en tres blocs: *(i)* abast i tipologia de la paràfrasi, *(ii)* creació i anotació de corpus de paràfrasis i *(iii)* la paràfrasi en la detecció automàtica de plagi.

En el primer bloc, partint de la base que els límits de la paràfrasi no són fixos, sinó que depenen de l'àrea de treball, la tasca i els objectius, es presenten tres casos límit de la paràfrasi: la pèrdua de contingut, el coneixement pragmàtic i la variació en determinats trets gramaticals. La caracterització de la paràfrasi pren una nova dimensió si l'observem des d'una perspectiva extensional. En aquesta línia, s'ha construït una tipologia general de la paràfrasi lingüísticament fonamentada. La tercera qüestió tractada en aquest bloc és la representació de la paràfrasi, essencial a l'hora de tractar-la formalment.

En el segon bloc, es presenta un mètode per a l'adquisició de paràfrasis relacionals a partir de la Wikipedia (*Wikipedia-based Relational Paraphrase Acquistion*, WRPA). Aquest mètode permet extreure automàticament de la Wikipedia paràfrasis que expressen una relació concreta. Utilitzant aquest mètode, s'ha creat el corpus WRPA, que cobreix diverses relacions i dues llengües (anglès i espanyol). Un subconjunt del corpus WRPA en espanyol i exemples extrets de dos corpus de paràfrasis en anglès s'han anotat amb els

tipus de paràfrasis que es proposen en aquesta tesi. Aquesta anotació ha estat validada aplicant les mesures d'acord entre anotadors (*Inter-annotator Agreement for Paraphrase-Type Annotation*, IAPTA), també desenvolupades en el marc d'aquesta tesi.

En el tercer i últim bloc, la tipologia proposada s'ha aplicat a l'àmbit de la detecció automàtica de plagi i s'ha demostrat que els tipus de paràfrasis més complexos i l'alta concentració de mecanismes de paràfrasi fan més difícil la detecció del plagi. També s'ha demostrat que les substitucions lèxiques i l'addició/eliminació de fragments de text són els mecanismes de paràfrasi més utilitzats en el plagi. Així, es demostra el potencial del coneixement parafràstic en la detecció automàtica de plagi i en la recerca en lingüística computacional en general.