Automatic recognition and extraction of multiword nominal expressions from corpora

Angeliki Fotopoulou, Giorgos Giannopoulos Maria Zourari, and Marianna Mini
Institute for Language and Speech Processing, "Athena" RIC
afotop@ilsp.athena-innovation.gr giann@imis.athena-innovation.gr

WP2

A. Introduction - Definition

This paper presents a first approach to the development of an algorithm that would automatically detect multiword expression (MWE). For the moment, our work focuses on nominal multiword expressions, since encoding their features according to grammar rules seems a rather feasible task.

A *nominal multiword expression* or a compound noun is a sequence of words that function as a noun, forming one lexical item. Most of the times, the meaning of the whole does not derive from the meaning of the parts. For example, *παιδική χαρά* [lit. "kids' joy", meaning "playground"], *ψήφος εμπιστοσύνης* ["vote of confidence"],

B. The algorithm for nominal MWE extraction

The algorithm that was developed for the present project is based in a combination of automatic MWE extraction methods (Sag et al., 2002). Combination of two methods:

- Word-based, knowledge-driven extraction: lexical sequences of a predetermined type are extracted (i.e., nominal compounds)
- Statistical extraction based on words: extraction of statistically idiosyncratic lexical sequences

Method:
(a) Six (6) general grammar rules are applied to a grammatically tagged corpus
(b) Results are filtered using more specific grammar rules and filters (word lists)
(c) Results are evaluated using statistical methods
(d) A linguist/encoder makes the final selection and MWE are stored in a Database

In order to test the performance of the algorithm, a corpus of about 142.000.000 words was used and nominal MWE of the following type were automatically detected:κράτος – μέλος (N + - + N) ["state member"], αγορά εργασίας (N + N(gen.)) ["labour market"], …

B.1. General/Initial Grammar Rules

i) *Adj* + *N* : (εκδοτικός οίκος ["publishing house"], μαύρη τρύπα ["black hole"])
ii) *N* + *N*_gen : (έργο τέχνης ["work of art"], σύνοδος κορυφής ["summit meeting"])
iii) *N* + *Article_def* + *N_gen:* (φαινόμενο του θερμοκηπίου ["greenhouse effect"],
iv) *N(-)N* : (κράτος-μέλος ["state-member"], λέξη-κλειδί ["key-word"])
v) [*Prep* + *N]* + *N:* (από μηχανής θεός ["deus ex machine"])
vi) *N* + *[Prep* + *N]:* (φόνος εκ προμελέτης ["premeditated murder"],

B.2. Filter Categories:

i) Rules to identify the cases where a MWE is likely to exist
ii) Rules to identify the cases where a nominal compound is unlikely to exist
In each category two sub-categories are identified:
i) General filters based on certain grammatical phenomena
ii) Filters of list of words or groups of words that are indicative or not of the presence of a nominal MWE

B.3 Statistical Processing

Log-likelihood Measure:

i) Widely used (WordSmith Tools 4, Collocate, Collocation Extract, Ngram Statistics Package)

ii) Better results than other methods we tried (high scores in more expressions accepted as multiword)

Finally, MWE are stored in classified structures (dbm hashes), along with information on the case used in each expression, its frequency, as well as the verbs it co-occurs with and their frequencies.

B.4. Final selection

Only expressions that are indeed nominal compounds are selected and information is added about: category of the compound (nominal, verbal, etc), register (general language, terminology, etc), type of the compound (fixed, collocation). The inflection of the parts of the expression is extracted from the morphological lexicon and its inflection is formed semi-manually. Finally, each expression accompanied with the above mentioned information is stored in the MWE Database.

C. Results

	Adj + N	**N + N(gen.)**	**N + N(gen.)**	**N + - + N**
1st Step: Application of 6 general rules	1.068.000	242.000	619.000	120.000
2nd Step: Application of specific rules filters	570.000	189.000	347.000	12.000
3rd Step: Application of statistical metrics	68.000	22.000	31.000	400

D. Future work

(a) Complete the collection of the material, (b)Refinement of the algorithm, (c) Test the algorithm on new written and spoken corpora, (d) Study and process 3-word nominal expressions (e) Study and extract adverbial and adjectival multi-word expressions

References

Anagnostou, N. & Weir, G. 2006. “Review of software applications for deriving collocations”. In *ICT in the Analysis, Teaching and Learning of Languages,* Preprints of the ICTATLL Workshop 2006, 21-22 Aug 2006, Glasgow, UK.

Evert, S. & Krenn, B. 2001. Methods for the Qualitative Evaluation ofLexical Association Measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France, 188–195.

Fellbaum, C. 1999. *La représentation des verbes dans le réseau sémantiqueWordNet. Langages*, (136):27–40.

Geyken, A. 2007. “The DWDS corpus: a reference corpus for the German
language of the twentieth century”, in C. Fellbaum (ed.), *Idioms and Collocations.*

continuum, London.

Kurz, D. & Xu, F. 2002. “Text Mining for the Extraction of Domain Relevant Terms and Term Collocations”. In *Proceedings of the InternationalWorkshop on "Computational Approaches to Collocations"*, Vienna.

Kyriakopoulou, T.; S. Mrabti; A. Yannacopoulou 2002. “Le dictionnaire électronique de mots composés en grec moderne” Lingvisticae Investigationes, John
Benjamins B.V, Amsterdam.

Manning, C. & Schütze H. 1999. “Foundations of statistical naturallanguage processing”. Cambridge, MA: MIT Press.

Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. 2002 “Multiword expressions: A pain in the neck for NLP”. In *A. Gelbukh, (ed.),Computational Linguistics and Intelligent Text Processing: Third InternationalConference*: CICLing-2002. Springer-Verlag, Heidelberg/Berlin.

Seretan, V., Nerima, L., & Wehrli, E. 2004. “A tool for Multi-word collocation extraction and visualization in Multilingual Corpora”. In *Proceedings ofthe Eleventh EURALEX International Congress (EURALEX 2004)*, 755-766,Lorient, France.