Extraction of Multilingual MWEs from Aligned Corpora

Eric Wehrli
LATL-CUI
University of Geneva

Aline Villavicencio
Institute of Informatics
Federal University of Rio Grande do Sul

WG2, WG3

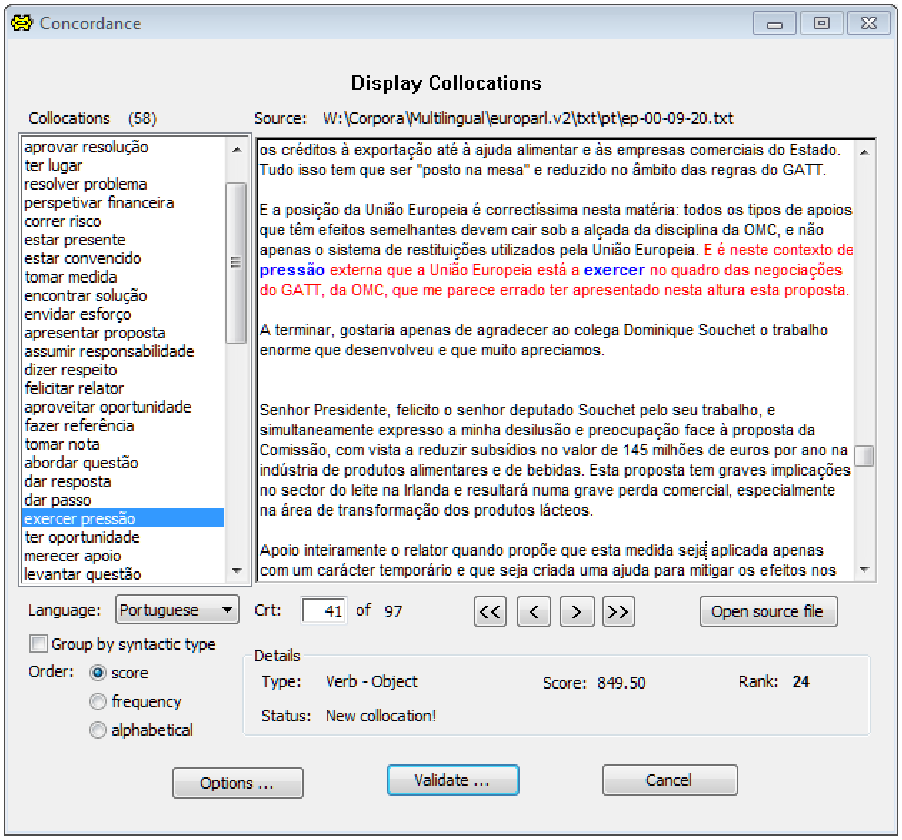
This paper presents an on-going research aiming at establishing a sizeable multilingual repository of MWEs extracted from aligned corpora. The languages involved are Portuguese, English and French, and the MWE types include collocations, named entities, compounds, idioms. In a first phase, MWEs will be extracted monolingually from the aligned sentences in Europarl corpus, and in a subsequent phase to other bilingual or multilingual corpora, such as the Brazilian newspaper Folha online.

MWE extraction involves two different methods : one, based on a symbolic parser, is used to extract collocations and other MWEs corresponding to grammatically well-formed phrases ; the second, based on statistical tools, is used to extract other types of MWEs, such as compounds (*by and large*, etc.), named entities and idioms. The goal of this work is to build the MWE resource in a way that increases coverage prioritizing precision, complementing the results of a precision oriented (parsing-based) with those from a recall oriented (statistical-based) method. In this paper we focus on the monolingual extractions, describing the two methods as applied to each language independently. In a multilingual phase we will adopt distributional methods to find MWE translations in the aligned sentences. The final MWEs lists will be made available to the community.

Collocation extraction

First, we use a symbolic parser to extract collocation candidates from pre-defined structural configurations (adjective-noun, noun-prep-noun, verb-direct object, and so on). Those candidates are then evaluated, using the log likelihood association measure, (see Seretan 2011 for a description of the whole system). The results are finally displayed in a specific user interface, along with the context in which they occur, as illustrated in figure 1 with Portuguese data. A lexicographer can then validate (or not) the collocations to be stored in the database. The parser as well as the collocation extraction tool work in the same way for the other two languages (French and English).

The parser used for this research is the Fips "deep" linguistic parser (Wehrli, 2007 ; Wehrli & Nerima, 2014), which computes constituent structures, as well as grammatical functions. In particular, it can cope with long-distance dependencies, such as the ones found in *wh*-questions and relative



Concordance

Display Collocations

Collocations (58)

aprovar resolução
ter lugar
resolver problema
perspetivar financeira
correr risco
estar presente
estar convencido
tomar medida
encontrar solução
envidar esforço
apresentar proposta
assumir responsabilidade
dizer respeito
felicitar relator
aproveitar oportunidade
fazer referência
tomar nota
abordar questão
dar resposta
dar passo
exercer pressão
ter oportunidade
merecer apoio
levantar questão

Source: W:\Corpora\Multilingual\europarl.v2\txt\pt\ep-00-09-20.txt

os créditos à exportação até à ajuda alimentar e às empresas comerciais do Estado. Tudo isso tem que ser "posto na mesa" e reduzido no âmbito das regras do GATT.

E a posição da União Europeia é correctíssima nesta matéria: todos os tipos de apoios que têm efeitos semelhantes devem cair sob a alçada da disciplina da OMC, e não apenas o sistema de restituições utilizados pela União Europeia. E é neste contexto de **pressão** externa que a União Europeia está a **exercer** no quadro das negociações do GATT, da OMC, que me parece errado ter apresentado nesta altura esta proposta.

A terminar, gostaria apenas de agradecer ao colega Dominique Souchet o trabalho enorme que desenvolveu e que muito apreciamos.

Senhor Presidente, felicito o senhor deputado Souchet pelo seu trabalho, e simultaneamente expresso a minha desilusão e preocupação face à proposta da Comissão, com vista a reduzir subsídios no valor de 145 milhões de euros por ano na indústria de produtos alimentares e de bebidas. Esta proposta tem graves implicações no sector do leite na Irlanda e resultará numa grave perda comercial, especialmente na área de transformação dos produtos lácteos.

Apoio inteiramente o relator quando propõe que esta medida seja aplicada apenas com um carácter temporário e que seja criada uma ajuda para mitigar os efeitos nos

Language: Portuguese

Crt: 41 of 97

□ Group by syntactic type

Order: ● score ○ frequency ○ alphabetical

Details

Type: Verb - Object

Score: 849.50

Rank: 24

Status: New collocation!

FIGURE 1 – User interface for collocation validation

clauses, as well as with constructions affecting the argument structure of verbs, such as passives and causatives [1].

Here are a few examples for the Portuguese collocation *tomar uma decisão*, with the French and English corresponding expression : *prendre une décision*, *to make a decision* : Notice that in all those examples, the two terms of the collocation occur in reverse order due to passive or relativization. This is precisely where the use of a linguistic parser comes handy. Generally speaking, we will show that extraction based on the parser is much less noisy and more accurate than the one based on purely statistical tools.

(1)a. As decisões são sempre tomadas pela própria Comissão.

b. Les décisions sont toujours prises par la Commission elle-même.

c. Decisions are always made by the Commission itself.

(2)a. Quaisquer decisões que se tomem...

b. toutes les décisions à prendre...

c. any decision to make...

(3)a. As decisões mais judiciosas são tomadas pelos consumidores europeus.

[1] The use of the Fips parser for the specific task of collocation extraction has been described in Wehrli et al., 2010.

b. Les décisions les plus judicieuses sont prises par les consommateurs européens.

c. Judicious decisions are made by European consumers.

Extraction of other MWEs

The method briefly outlined above gives very precise results for collocations and other grammatically well-formed expressions. However, it depends on parsing coverage for a given language and there are other types of MWEs for which it cannot be used. For instance, it cannot easily retrieve compounds which are not grammatically well-formed, such as by and large, or French donnant donnant. Even more important, it doesn't work well at all with named entities. For such MWEs, we use the mwe-toolkit (Ramisch, 2015) based on the definition of appropriate linguistic (POS tag) patterns to extract an initial list of candidates and the use of statistical association measures, to identify candidates whose components co-occur more often than chance. In a first stage this extraction is done monolingually, independently for each language. An initial set of candidates is obtained for each language, and is automatically validated according to the agreement between the syntactic and statistical-based methods, and for the cases where they disagree, using a weighted voting scheme based on how confident they are about each candidate. In a final stage we will apply alignment and distributional methods to identify bilingual correspondences for the monolingual candidates (Caseli et al. 2010, Laranjeira et al. 2014).

Bibliographie

Caseli, H.M., M.G.V. Nunes, C. Ramisch, A. Villavicencio (2010). "Alignment-based extraction of multiword expressions" *Language Resources and Evaluation, Special Issue on Multiword Expressions*, Volume 44, Number 1-2, p. 59-77,

Laranjeira, B., V. Moreira, A. Villavicencio, C. Ramisch, and M.J. Bocorny Finatto (2014). "Comparing the quality of focused crawlers and of the linguistic resources obtained from them", *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), Reykjavik, Iceland.

Ramisch, C. 2015. *Multiword Expressions Acquisition. A Generic and Open Framework*, Springer Verlag.

Seretan, V. 2011. *Syntax-Based Collocation Extraction*, Springer Verlag.

Wehrli, E. 2007. "Fips, a 'Deep' Linguistic Multilingual Parser" in *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, pp. 120-127, Prague, Czech Republic.

Wehrli, E. & L. Nerima, 2014. "The Fips Multilingual Parser", in N. Gala, R. Rapp, and G. Bel-Enguix (eds.), *Language Production, Cognition and The Lexicon.* Text, Speech and Language Technology 48, Springer. pp. 473-490.

Wehrli, E., V. Seretan & L. Nerima, 2010. "Sentence analysis and collocation identification", in sl Proceedings of the Workshop on Multiword Expressions : from Theory to Applications (MWE 2010), pages 27–35, Beijing, China.