PARSEME shared task on automatic detection of verbal MWEs

Struga
7 April 2016

Introduction

- Wide range of researchers interested in MWEs and parsing
- Several corpora and tools are available
- It is problematic to directly compare results obtained on different datasets and/or different methods
- Shared task with standardized annotation principles, corpora and evaluation metrics

The shared task: aims and goals

- As multilingual as possible
- Focused scope: verbal MWEs
- Task: identify their occurrences in running text
- Standardized categories and annotation principles
- Evaluation: standardized evaluation metrics

Progress report

What has happened since the Iași meeting?

- Organization and management
- Datasets
- Annotation principles
- Evaluation methodology
- Challenging examples
- Annotation tools

Organization and management

4 language groups containing 19 languages

Organizers: Agata Savary, Antoine Doucet, Veronika Vincze

Technical support: Federico Sangati, Behrang QasemiZadeh

- **Germanic languages** – group leader: Fabienne Cap
  - English (Corina Forascu et al., Alessandro Lenci, Ismail El Maarouf, Federico Sangati et al., Veronika Vincze et al.)
  - German (Fabienne Cap, Agata Savary)
  - Swedish (Fabienne Cap et al.)
  - Yiddish (Yaakov Ha-Cohen Kerner, Chaya Liebeskind)
- **Romance languages** – group leaders: Marie Candito and Carlos Ramisch
  - French (Marie Candito, Matthieu Constant, Ismail El Maarouf, Yannick Parmentier, Carlos Ramisch, Agata Savary)
  - Italian (Alessandro Lenci, Johanna Monti, Federico Sangati, Simonetta Vietri)
  - Romanian (Corina Forascu et al., Verginica Mititelu)
  - Spanish (Carlos Herrero, Carla Parra)
  - Brazilian Portuguese (Aline Villavicencio, Carlos Ramisch, Leonardo Zilio)

Organization and management

- **Slavic languages** – group leader: Ivelina Stoyanova
  - Bulgarian (Ivelina Stoyanova, Svetla Koeva, Tsvetana Dimitrova, et al.)
  - Croatian (Marko Tadić et al.)
  - Polish (Monika Czerepowicka et al.)
  - Slovene (Simon Krek et al.)
- **Other languages** – group leader: Voula Giouli
  - Farsi (Behrang QasemiZadeh, Mojgan Seraji)
  - Greek (Voula Giouli et al.)
  - Hebrew (Yaakov Ha-Cohen Kerner, Chaya Liebeskind)
  - Hungarian (Veronika Vincze et al.)
  - Maltese (Lonneke van der Plaas, Mike Rosner)
  - Turkish (Gülşen Eryiğit et al.)

Annotated datasets & guidelines

- Data with MWE annotation on the basis of standardized (universal) guidelines
- (Re)annotation efforts are required
- Possibly texts from newspapers / Wikipedia (to be discussed)
- Corpus size: 3500-4000 MWE occurrences per language (approx. 18-20K sentences, based on earlier annotation for English)
- Annotation guidelines written in English
- Harmonizing theoretical and computational linguistic considerations
- Basic principles:
  - Each verbal MWE occurrence is annotated
  - Subcategories are annotated
  - Non-contiguous elements are also annotated
- To be adapted to the given language by the annotator team
- A part of the data should be double-annotated to check IAA
- As of April 2016: two pilot annotation phases over

Basic definitions

Multi-token word (MTW): contains several tokens, e.g. *Pandora's*, (PL) *SMS-ować* 'to write an SMS'
Multi-word token (MWT): contains several words, e.g. (IT) *della* = *de la*, (FR) *court-circuiter* 'to short circuit'
Multi-Word Expressions (MWEs): (continuous or discontinuous) sequences of words which:

- contain at least **two lexicalized words**, including a head word and at least one other syntactically related word
- show some degree of morphological, syntactic and semantic **non-compositionality**

Collocations, i.e. word co-occurrences whose idiosyncrasy is of statistical nature only (e.g. *the graphic shows*, *drastically drop*, etc.) are disregarded.

Verbal MWEs

Verbal MWEs include three syntactic types:

- **Prototypical verbal MWEs** function as (possibly unsaturated) verb phrases, e.g. *made the final decision*, *will break her heart*, *took this to heart*
- **Nominal and participial variants** of prototypical VMWEs maintaining their idiomatic reading, e.g. *decisions which we made*, *decision making*, *heart-breaking*
- **Sentential MWEs**, e.g. *a little bird told sb*, *the problem lies in sth*, *the early bird catches the worm*, *better late than never*

Non-compositionality tests

At least one of the following should hold for VMWEs:

- Presence of a cranberry word, e.g. *to go astray*
- Lexical inflexibility, e.g. *#to allow the feline out of the container* (to let the cat out of the bag)
- Morphological inflexibility, e.g. *#to take a turn* (to take turns)
- Morpho-syntactic inflexibility, e.g. *#I give you his word for that* (I give you my word for that)
- Syntactic inflexibility, e.g. *#Bananas are gone* (to go bananas)
- Semantic non-compositionality, e.g. *to kick the bucket* = to die

VMWE typology

2 universal categories, i.e. valid for all languages participating in the task:

- light verb constructions, e.g. *to give a lecture*
- idioms, e.g. *to go bananas*

3 quasiВн-universal categories, valid for some language groups or languages but not all:

- verb-particle constructions, e.g. *to do in*
- inherently pronominal verbs, e.g. (FR) *se suicider* 'to suicide'
- inherently prepositional verbs, e.g. *to come across sth*, *to rely on sth*

languageВн-specific categories, to be defined for each language concerned

other verbal MWEs, which gather the types not belonging to any of the categories above e.g. *drink and drive*, *fortune favors the bold*, *better late than never*.

Light verb constructions

(1) The candidate consists of a verb and a noun which is predicative.
(2) The noun has one of its usual meanings.
(3) An NP headed by the noun can be formed, containing all the syntactic complements of the verb, and such an NP denotes the same event or state as the one denoted by the LVC candidate.
Paul had a nice walk – Paul's nice walk / the nice walk of Paul
(4) One syntactic argument of the verb is a semantic argument of the noun, which can be tested by the impossibility to realize such an argument twice.
*Paul made a decision – *Paul made the decision of the committee*
Paul leads the discussion. – Paul leads the discussion of the committee
(5) The verb is not used in one of its original sense(s), e.g. in *to deliver a speech*, a speech is not literally moved to another place.
(6) If the predicative noun is the verb's direct object, it must be possible to passivize the construction.

Idioms

Idioms comprise a head verb and at least one lexicalized argument; the latter assumes any function (i.e. subject, direct or indirect object, prepositional complement, or any combination thereof)
Idiomaticity is attributed primarily to the fact that they have a non-compositional meaning.
Tests have been created to distinguish idioms and other VMWEs.
kick the bucket
spill the beans

Verb-particle combinations

(1) They are formed by a head verb and a particle.
(2) Both the verb and the particle are lexicalized.
(3) The meaning of the VPC is non-compositional. Notably, the change in the meaning of the verb goes significantly beyond adding the meaning of the particle (e.g. to do in = to kill).
(4) The verb and the particle can sometimes be separated with a noun or pronoun without any change in meaning (e.g. *spit (it) out*).
(5) There is often an English synonym or a translational equivalent in another language which is a one-word unit or is a verb with a verbal prefix (e.g. *get away* and *escape*).
(6) A noun can often be derived from it (e.g. *breakthrough*).
(7) Multi-word tokens should also be annotated (e.g. (GER) *aufgepasst* – past participle form of 'to pay attention').

Inherently pronominal verbs

Inherently pronominal verbs are full verbs combined with a reflexive clitic (REFL) *self* where:
(a) the verb never occurs without the clitic, e.g. (FR) *se suicider*to suicide
(b) the clitic markedly changes the meaning of the verb, e.g. (FR) *s'apercevoir* 'to realize' vs. *apercevoir* 'to perceive'
They are very common in Romance and Slavic languages.
Some tests to annotate them:

1. The verb only exists with the REFL and never occurs without it
   (SP) *suicidarse* 'suicide' – **suicidar*
2. Given the same verb without the REFL, if all of its meanings are clearly different from the pronominal form, then it is an IPronV.
   (FR) *s'agir* 'to be (a matter of)' – *agir* 'to execute an action'
3. Given the same verb without the REFL and with the same/similar meaning as the pronominal form, if it has a different subcategorization frame, then it is an IPronV.
   (PT) *esquecer* 'to forget' takes a direct object – *se esquecer* requires an indirect prepositional phrase with *de* 'of'

Inherently prepositional verbs

Verb + preposition combinations where:
(i) the verb mandatorily requires a preposition but the meaning of the verb is more or less transparent (e.g. *refer to*)
(ii) adding the preposition markedly changes the meaning of the verb (e.g. *come across*)
They are common in English.
Some tests (based on DIMSUM):
The object and the preposition cannot be separated by the prepositional object (*He depended on me* vs. **He depended me on*).
A circumstantial PP be inserted between the verb and the preposition (*I could rely on him at once - I could rely at once on him*).
The verb usually does not occur without its prepositional complement (**He referred*).

Other VMWEs

VMWEs which do not fit any of the preceding categories, including notably:

- verbal expressions with **no lexicalized arguments** such as *to drink and drive*, *to tumbleВн dry*
- **proverbs**, i.e. sentences expressing facts thought to be true by most people, e.g. *Fortune favors the bold*, possibly with omitted head verbs, e.g. *better late than never*, (FR) *loin des yeux, loin du coeur* 'far from the eyes, far from the heart'
- totally lexicalized and often morphologically and syntactically **frozen phrases**, e.g. *the pleasure is mine*, (PL) *I tu jest pies pogrzebany* lit. and here is the dog buried 'here is the essence of the problem'
- **exclamations**, e.g. *I beg you pardon!*, (PL) *Co ja widzДЄ!* lit. what do I see?! 'what a surprise!'
- **similes**, e.g. *to sleep like a log*

Annotation methodology

- Some of the texts are annotated by more than one annotator
- Annotators are not allowed to discuss or share commonly annotated data with each other
- Annotations can be compared
- Inter-annotator agreement rate can be measured
- Problematic issues can be discovered in this way
- Annotators can signal their confidence (i.e. whether they are sure the given sequence belongs to a specific MWE type)

Evaluation methodology

Inter-annotator agreements (IAAs): to assess the performance of annotators in the processes of
(a) identifying text boundaries of VMWEs in text – F-score
(b) assigning the identified VMWEs to one of the categories discussed – Cohen's κ measure

Language	Token#	A1	A2	F-Score	κ
English	5,725	0.148	0.233	0.407	0.616
Farsi	4,913	1.63	1.335	**0.708**	0.639
French	4,218	0.26	0.215	0.568	0.076
German	4,266	0.151	0.435	0.555	0.8
Greek	6,313	1.213	1.153	0.638	0.428
Hungarian	4,607	0.456	0.118	0.325	**0.914**
Italian	3,808	0.105	0.155	0.269	0.222
Polish	3,397	0.133	0.197	0.299	0.333
Portuguese	4,127	0.13	0.14	0.296	0.385
Romanian	4,870	0.25	0.24	**0.735**	**1**
Spanish	6,112	0.854	0.864	0.473	0.172

Annotation tool & technical issues

Friday's session will be dedicated to these issues.

Questions to discuss in Struga

- choice of the tool (Friday)
- choice of the corpus genre
- availability of the corpus (open availability vs. compatibility with the existing corpora)
- substantial changes to the guidelines if needed, based on challenging examples
- language-specific examples not compulsory for the whole guidelines

Challenging examples from Slavic languages

ambiguity between a VMWE and an embedded MWE
РўРчРџРўРөРё [СЛЪРҳСҐР„Р« Р„Рө РҫСЅРћРњР«] - see [black on white]
РўРчРџРўРөРё [Рў Р„Р«РўРө СҒРўРҳСЂРњРчР„Рө] - see [in a new light]
РәСЂР»СЂРўРөРё [Р„Рө РҳРўСҐР«] - buy [in bulk]
The examples above show some restrictions on the verb (e.g., combine with vizhdam, in metaphoric meaning, but not with other mental perception verbs), but the verb is in one of its regular meanings and there is no single verb with the meaning of the whole vMWE, hence the hesitation.

Challenging examples from Slavic languages

give a different (than OTH) label in Bulgarian to MWEs which stem from verbs but lost their verbal meaning
СҐРөРчРҫРҷСҐРө/understand СҒРҳ/REFL.PASSIVE - (it) is understood (of course)
РәРөРәСЂР«/whatever Рҷ/and РұРө/to Рҳ/be - whatever to be (whatever)
Р„Рҳ/not СҚРҳСЇ/want РњРҷ/PARTICLE - you not wanting (unexpectedly)
a language specific label as they are common in Bulgarian.