STSM PARSEME
Université Aix-Marseille
LIF (Laboratoire d'Informatique Fondamentale)
4th-8th July 2016
Invited researcher: Beatriz Sánchez Cárdenas
Host: Carlos Ramisch

Final Report

A. INTRODUCTION

The main goal of this STSM was to use the MWEtoolkit[1] developed in Marseille in order to automatize and guide the creation and enrichment of semantic frames in Ecolexicon, developed in Granada. More specifically, one of our aims was to develop efficient semi-automatic strategies to extract phraseological information from the corpora for terminological and translation purposes and, secondly, the representation of this information in terminological resources.

Preparatory work implied assembling the corpora and tokenizing it with a POS tagger in order to be able to analyze it with MWEtoolkit. Our corpora is composed of three different subdomains of Environmental Sciences (Extreme Events, Wastewater Treatment and Erosion) and three languages (English, French and Spanish), each of these six corpora has at least one million tokens. The questions that we wanted to find an answer for during this research stay are described below.

Question I

Finding the equivalent verb associated to a term in a pararel corpus. Obviously, bilingual dictionaries cannot solve this problem. Let's observe this example:

- The volcano **worked** in 1857 last time, before the eruption in 1987.

The equivalent verb in Spanish is not the equivalent verb in the general language, "trabajar", but "**estar activo" (be active)**, which of course does not appear in any

1 http://mwetoolkit.sf.net

bilingual dictionary. The only possible way to find these kind of equivalences is looking for the same actantial structure in a parallel corpus.

Question II

We need to implement a methodology to identify which are verbs associated to each term in the corpora. Sketch Engine (SkE) offers some interesting results using CQL such as: [lemma="volcano"] []{0,3} [POS=V.*]
Nevertheless, results obtained with SKE have some drawbacks: there is lack of pertinence (verbs like *to do, to be, to say*), they have a high degree of redundancy and they cannot be exported to an xml file. It might be interesting to compare the results obtained with SkE and with MWEtoolkit.

Question III

Verbal hierarchies in specialized domains (Sánchez & Buendía 2012) done "manually" might not be accurate enough and, most importantly, they might not reflect the language that is actually being used by specialists. We compare the results obtained manually and automatically in order to figure out what kind of methodology is more accurate for automatic translation prediction.

B. WORK FLOW DESCRIPTION

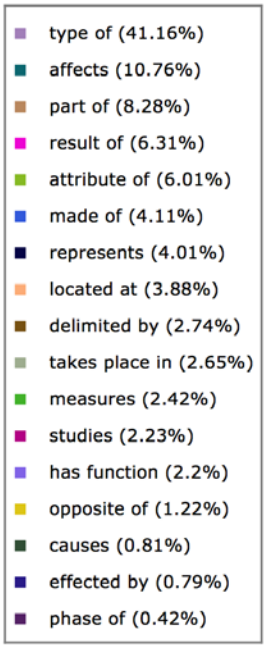
In order to developed a suitable methodology to extract meaningful noun-verb combinations form the corpora, we did a pilot experiment with the restricted subdomain of Volcanology. Our main goal was to establish a process that could be lately implemented with larger corpora in other domains and languages.

1. Pilot study with a restrained domain

Creating queries with MWEtoolkit to retrieve verb-noun combinations

Our goal was to retrieve from the corpora all the verbs that lexicalize a specific semantic relation between two concepts in Ecolexicon ontology. For instance the relation VOLCANO VERB LAVA, lexicalized by verbs such as: *expels, erupts, ejects*.

We used EcoLexicon semantic relations between concepts as a starting point:



In the first place, it was necessary to define a query that would retrieve the appropriate lexical items from the corpora. To this end, we chose a concept of the specialized domain of Extreme Events, namely VOLCANO, as well as the concepts semantically related to it, such as the concept LAVA. Secondly, we looked in Ecolexicon database for the terms that lexicalize this relation. Then we looked in the corpora for other terms lexicalizing this relation. We enlarged the results looking in the corpora for other terms sharing the conceptual structure "CONCEPT1 VERB CONCEPT1". In other words, combining Ecolexicon semantic relations with the most typical nouns in the context of those concepts, we can retrieve patterns that might be used to interrogate the corpus. The final result is the definition of a search query that will allow us to extract all the verbs that lexicalize a given semantic relation.
We performed the same types of queries using both the mwetoolkit and SkE in order to later compare the results.
This strategy is similar to bootstrapping, when one departs from minimal seeds and then builds on the system's output itself in order to generalize and obtain new data.

2. Extraction of verb-noun combinations from the corpora

Defining MWE queries

We query the corpus using the mwetoolkit. Therefore, we performed the following preparatory steps:

1. **Gather the corpora:** this was preliminary work in the Ecolexicon project. The corpora include manually selected documents and web-crawled texts using webBootCat and specific seed keywords.
2. **POS-tag the corpora:** this was performed by importing the raw text on Sketch Engine and then applying one of the available taggers and lemmatizers: TreeTagger for English and French and Freeling for Spanish. Further preprocessing in the form of depencendy trees is being currently investigated using UDPipe.
3. **Index the corpora**: the mwetoolkit can deal with very large text collections, but a preliminary indexation must be performed in order to create data structures

which allow querying and counting word sequences in the text. We have created binary indices for all corpora and for all available information: surface forms, lemmas and POS tags.

4. **Query the corpora**: corpus queries take the form of special multi-level regular expressions. We used the textual patterns in order to match combinations including the target terms. Once we have tuned the queries, we encapsulated them into easy-to-use scripts which abstract away some gory details of regular expressions, increasing the readability.
5. **Filter and sort the output**: one of the main features of the mwetoolkit is that it includes many filters that help cleaning query results. This is essential because, in a larger-scale experiment, it helps speeding up lexicographic work by looking only at relevant output. We use a simple association measure, pointwise mutual information (PMI), to sort the query results in descending order.

For example, in order to look for the verbs that lexicalize the relation between VOLCANO and LAVA we made this query:

```
../bin/candidates.py -f -e '[lemma="volcano"] []{repeat={0,3} ignore=true} [pos~/V.*/] []{repeat={0,3} ignore=true} [lemma="lava"]   ' -g ExtremeEvents.info >CANDIDATES.xml

'[lemma="volcano"] []{repeat={0,3} ignore=true} [pos~/V.*/] []{repeat={0,3} ignore=true} [lemma="lava"]   '
```

Since the mwetoolkit commands tend to be complex and cannot be memorized, we developed some scripts in order to accelerate the work:

- **./index-all-corpora.sh**

As the name indicates, indexes all corpora at once. It only needs to be run if new copora are added or if corpora are modified (e.g. deduplication).

- **./search-pattern.sh <pattern> <corpus-folder>**

Looks for the regular-expression pattern in the indexed corpus contained in the corpus folder. Afterwards, counts the output and individual words and calculates the association measur, sorting the output in descending order. Result is a list in TSV (tab-separated values) format, editable in Excel, named search-result.tsv. This file must be renamed to save the query result, otherwise it is overwritten.

- **./search-triples.sh <noun1> <verb> <noun2>**

This script uses the previous one in order to look for patterns corresponding to a noun (noun1), followed by a verb, followed by another noun (noun2), but allowing 0 to 3 intervening words to appear in between the verb and each noun. These words are then discarded from the output, since they will often correspond to adverbs, determiners and prepositions that do not carry much semantic information for this task. Each of the three elements can be underspecified by using the special keyword “ANY”, which means that the query will return any nominal or verbal lemma in that position. For example, the query “ANY” “ANY” “ANY” would match all noun-verb-noun pairs where there are no more than 3 intervening words between the verb and each noun. This script also deals with the fact that POS tagsets are different in the three target languages, by switching the nominal and verbal tags according to the language code prefixing the corpus folder name.

Using Ecolexicon semantic relations to define MWE queries

a) From N-N relations to V

I) VOLCANO [CAUSE] N

Taking as a starting point the CAUSE relation between the concepts of VOLCANO and ERUPTION, we formulated this query:

```
./search-triples.sh volcano ANY eruption EN_Extreme_events_anotated/
```

The above query retrieved the verbs *create, produce*. Using a bottom-up approach, these verbs were reused to look up for new occurrences of the structure "volcano causes something" with the search:

```
./search-triples.sh volcano "(cause|produce)" ANY EN_Extreme_events_anotated/
```

This query retrieved N such as *continent, land masses, explosion, flow, destruction, death*. They were used for a new query:

```
./search-triples.sh volcano ANY  "(eruption|continent|land|explosion|flow)"
EN_Extreme_events_anotated/
```

(Results in xml file "EN_ExtremeEvents_Results_toolkit-SKE)

The information retrieved from the previous queries shows different patterns. It becomes clear that the combination of VOLCANO with different verbs and nouns lexicalizes different semantic frames.

II) VOLCANO [PART OF] N

In EcoLexicon, the concept VOLCANO has a PART OF relation with the concept LAVA, since this geological material (lava) is conceptualized as one of the components of a volcano.

Using the procedure described above, we obtained other terms with the same PART OF relation, such as *gaz, smoke, ash, cloud, lava, rock, material, dust, steam.* We used these results to formulate this query:

```
./search-triples.sh "(gaz|smoke|ash|cloud|lava|rock|material|dust|steam)" ANY
volcano
```

(Results in xml file "EN_ExtremeEvents_Results_toolkit-SKE)

III) VOLCANO [LOCATED AT] N

This query, allowed us to know where are *volcanos* typically located:

```
./search-triples.sh volcano ANY  "(island|continent|land)"
EN_Extreme_events_anotated/
```

The results analysis showed that the POS tag for “proper noun” had not been so far included in the triples script, which is an important piece of information since most geographical accidents are located in places that have a proper name. This observation gave us the occasion to improve the script adding an option that allows to easily include or exclude proper names from the output of the queries, independently of the language.

b) From V to N-N relations

Using the verbs obtained in the section previously described, we were able to extract new N-N dependencies from the corpora.

```
./search-triples.sh ANY
“(erupt|cause|create|shift|change|move|destroy|make|form|emit|spew|blech|emit|prod
ANY EN_Extreme_events_anotated/
```

This is a valuable information since it shows the whole pattern of the verb-noun combinations of *volcano*. We can classify these verbs according to the noun of the complement.

3. The participants (frame elements) of VOLCANO

The information described above was used to extract all the semantic participants to the frame of volcano. In Framenet semantics, these semantic participants are called “frame elements”. In EcoLexicon, we make a difference between the semantic roles, also known as thematic relations (such as Agent, Theme, Experiencer) in other Linguistics traditions, and the linguistic realizations instantiating all the semantic categories, which could be compared to a noun typology (such as NATURAL DISASTER, ATMOSPHERIC AGENT, HUMAN BEING).

Thematic relations

Agent: *volcano, explosion, eruption*

Theme: *gaz, smoke, ash, cloud, lava, rock*

Result: *island, land, deaths, island*

Patient: *people, houses, city, coast,*

Semantic categories
LANDFORM: *continent, land, island, volcano*
MATERIAL ENTITY: *gaz, smoke, ash, cloud, lava, rock*
NATURAL DISASTER: *eruption, explosion*
DAMAGE: *death, loss of property*

4. The semantic frames of volcano

Frame 1: cause_damage

Definition: An **Agent** (LANDFORM | MATERIAL ENTITY | NATURAL DISASTER) causes a negative **Result** (DAMAGE, DEATH, LOSS OF PROPERTY) negatively affecting a **Patient** (AREA, HUMAN BEING).

Linguistic realizations of frame elements:

Agent: *volcano, ash, volcanic eruption*
Result : *death, killed people, damage, injuries, problems, island, continent*
Patient : *people, coast, homes*
Verbs: *cause, produce, kill, contaminate, damage*

Examples:

People have died from **volcanic blasts.**
The **death toll** from Japan's **volcanic eruption** has risen to 47 after more **victims** were discovered on the ash-covered summit.
The **ash can** also kill **plants**, contaminate **water supplies** and damage **electronic equipment.**
Since **Turrialba Volcano** re-awoke last October, **volcanic ash** has dirtied **homes,** damaged **crops** and mucked up **travel plans**.
Ash can cause **respiratory problems, throat problems** and **burning** in the **eyes** or **skin**.

cause_damage		
frame element	**semantic class**	**linguistic realization**
Agent	LANDFORM	*volcano*
	MATERIAL ENTITY	*ash*
	NATURAL DISASTER	*volcanic eruption*
Result	DAMAGE	*death, killed people, damage, injuries, problems, island, continent*
Patient	AREA	*coast, homes*
	HUMAN BEING	*people*

Frame 2: cause_motion

Definition: An **Agent** (LANDFORM | NATURAL DISASTER) ejects a **Theme** (MATERIAL ENTITY) with sudden force.

Frame elements:
Agent: *volcano, volcanic eruption*
Theme: *gaz, ash, lava, rock, dust*
Verbs: *flows, eject, spew, throw*

Examples:
Mount Merapi, Indonesia's most volatile **volcano** , spews **clouds of ash** over the island of Java on Monday.

Magma that flows out of a **volcano is** called **lava.**

When the **lava is** thrown from the **volcano** , it solidifies into small particles.

In February, for instance, the restless **volcano** spewed **clouds of ash** 2 kilometers into the sky, said Indonesia's disaster management agency.

The **volcano** can easily eject dangerous **ash** to aircraft cruise altitudes, and disperse it over large areas.

cause_motion		
frame element	**semantic class**	**linguistic realization**
Agent	LANDFORM	*volcano*
	NATURAL DISASTER	*volcanic eruption*
Theme	MATERIAL ENTITY	*gaz, smoke, ash, cloud, lava, rock*

Frame 3: cause_existence

Definition: An **Agent** ejects a **Theme** (MATERIAL ENTITY) provoking a **Result**.

Agent: *volcano, volcanic eruption*
Result: *continent, landmass, island*
Verbs: shift, create, make, become

cause_existence		
frame element	**semantic class**	**linguistic realization**
Agent	LANDFORM	*volcano*
	NATURAL DISASTER	*volcanic eruption*
Result	LANDFORM	*continent, landmass, island*

Examples:

I cannot easily imagine volcanoes shifting continents and creating landmasses.

The volcanoes allow the continents to move because it heats up the rock.

When volcanoes erupt, they create massive earthquakes.

Volcanic eruptions may have been a small factor for creating land.

This massively violent eruption can definitely shift an entire continent.

Many islands were created by volcanoes erupting at sea, and many more will be made in the future. Although I have never been there, I am aware that the islands of Hawaii were made entirely by volcanoes.

If the volcano erupts in the middle of the ocean, the rocks that have been created by the volcano will make entirely new landmasses.

Hot lava erupts from a volcano and when it cools it becomes new land.

5. Implementation of the methodology in other domains and languages

We are currently working on this.

C. RESULTS AND PERSPECTIVES OF THIS STSM

1. We have sent a contribution to the FrameNet Terminology Workshop at ICCG9. It has been accepted: http://www.ufjf.br/iccg9/home/theme-sessions/frame-based-accounts-of-specialist-languages/
2. We will make a contribution to a special number of IJL whose proposal has already been sent to the publisher.
3. We will write an article to a JCR journal such as Terminology.
4. We will populate EcoLexicon with these results in view of CAT tool.

The work carried out during the STSM carried out by Beatriz Sánchez Cárdenas is hereby confirmed by the host, Carlos Ramisch, and conforms to our common research interests and goals. We will pursue this collaboration remotely via regular Skype meetings and try to meet again personally soon.