T_2O — Recycling Thesauri into a Multilingual Ontology

José João Almeida, Alberto Simões

Departamento de Informática
Universidade do Minho
Braga, Portugal
{jj, ambs}@di.uminho.pt

Abstract

In this article we present T_2O – a workbench to assist the process of translating heterogeneous resources into ontologies, to enrich and add multilingual information, to help programming with them, and to support ontology publishing. T_2O is an ontology algebra.

1. Introduction

Dictionaries and Thesauri are valuable resources for Natural Language Processing but do not exist as freely available as expected, especially for languages other than English and, when they exist, they are just available for querying on-line.
Our main goal with T_2O[1] is to create a multilingual ontology:

- freely available on-line and to download;
- with a computer readable format;
- with a good API;
- with a structure as rich as possible;
- reusing all the structured information we can get;

so that it can be used for computational-linguistic tasks, such as machine translation, named entity recognition or information retrieval.
T_2O aims not just to produce these resources but prepare a set of tools to bootstrap other ontologies, to enrich and to publish them. These resources will be made available as open-source software.

1.1. Dictionaries vs. Ontologies

A Dictionary is a word-oriented view over a complex semantic network of concepts. Traditional dictionaries connect single word entries with associated information, and normally they have a *multi-senses* structure to deal with ambiguity of words.
Ontologies, thesauri and terminologies are focused in concepts and semantic relations between concepts.
In order to focus in semantic relations, it is crucial to reduce as much as possible term ambiguities. The main mechanisms used to reduce this ambiguity are:

- the use of multi-word terms: defining a context reduces ambiguity;
- definition of a domain for each concept: some of the ambiguities disappear if the domain is known. (e.g. Turkey in the domain of geography is different from the turkey in the animal domain);
- use a set of terms to represent each concept;
- for each concept, choose a preferential term (with low ambiguity) and a set of cross-reference terms.

We are aware of the fact that term ambiguities exist, and we are trying reduce it, but we could not find any good way of dealing with large scale sense-disambiguation.

2. T_2O = Ontology Algebra

To define an algebra we need to define sorts and operations. While the main sort will be *ontologies*, other will appear as well.
Our **initial resources** are:

- some thesauri available on the WEB (e.g. the Thesaurus of UNESCO (UNESCO, 1995), the EuroVoc Thesaurus (EuroVoc, 2004), the TEE (Community, 1991)). This kind of resources have a *good* structure, but sometimes are not available in Portuguese, or the number of terms is not very big;
- terminologies, glossaries and vocabularies;
- lists of some classes of elements (e.g. the list of the birds of Spain; the list of the countries and their capitals);

In order to take advantage of a so heterogeneous set of resources, we need a **toolkit** or **ontology algebra** with:

- a common type-system (sorts are: ontologies, tables, terms, information about the terms);
- a set of tools to transform the resources available in sorts of the previous type-system;
- a set of tools to conciliate and join the basic sorts;
- a set of tools to transform the basic sorts in sorts appropriated for the external user needs (to create several views over the information);
- a set of tools to create new thesauri (e.g. the domain specific language **tabularThesaurus** presented on subsection 4.);
- a set of tools to perform rule-based forward chaining completion;

[1] T_2O — Thesaurus to Ontology Workbench

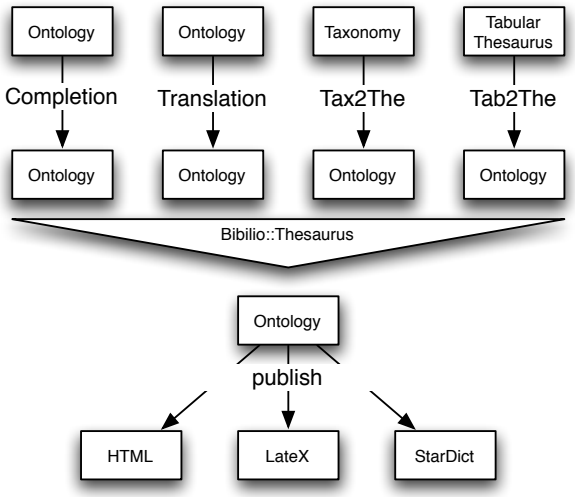


Figure 1: Ontology Algebra

Many of the tasks related to the ontology construction are transformations and unions of information: term conciliation, ontology completions, ontology conciliation, translation of ontologies, inversion of ontologies to another language, transformation of tables into ontologies, transformation of taxonomies into ontologies, and so on.

2.1. Ontology Definition

To manage ontologies we use a Perl module named `Biblio::Thesaurus`[2] (Simões and Almeida, 2002) that manages ISO-like monolingual and multilingual thesauri (ISO 2788, 2002; ISO 5964, 2002). In fact, the module manage much more than simple and traditional thesauri, but objects very similar to ontologies:

$$\begin{array}{rrcl} \mathcal{T} = term \mapsto (& \mathcal{L} & \mapsto & term_{\mathcal{L}} \quad \times \\ & rel & \mapsto & term^{\star} \quad \times \\ & extrel & \mapsto & any \quad) \end{array}$$

Explaining the math, in our ontology we relate each term in a base language to information: translations of that term in other languages ($\mathcal{L} \mapsto term_{\mathcal{L}}$), other terms related with the current one ($rel \mapsto term^{\star}$) and other kind of information associated to the term ($extrel \mapsto any$).
The relations are the ones defined in (ISO 2788, 2002): BT: broader term; NT: narrower term; UF: use for; USES; RT: related term; SN: scope note; but are not restricted to this set. `Biblio::Thesaurus` copes with any other relation you might define.
Notice that the *scope note* is an external relation because it associates assorted text with the term.
`Biblio::Thesaurus` is a flexible tool and allows the users to define for each relation its range: if they are normal relations (to another term), external relations (like the scope note, where we associate text with the term) or languages.

[2] While the name of the module refers to thesaurus, it is powerful enough to manage ontologies. The name was maintained just for historic reasons. This also explains why we use interchangeably thesaurus and ontology in this article.

In the ISO Thesaurus syntax tradition, `Biblio::Thesaurus` thesauri are defined in plain text files with concepts separated by empty lines. Each concept has its representative term in the first line. Following lines contain the relation and the term or list of terms related. See following example to get a better idea of how they work.

```
Acid rain
FR Pluie acide
ES Lluvia ácida
PT Chuva ácida
MT Pollution, disasters and safety
BT Air pollution
RT Pollutants
RT Precipitation
RT Rain

Pollutants
FR Polluant
ES Contaminante
PT Poluentes
MT Pollution, disasters and safety
NT Wastes
RT Acid rain, Bacteria, Carbon dioxide
RT Chemicals, Dangerous materials
RT Fertilizers, Hydrocarbons
RT Pesticides, Petroleum
```

In addition, the `Biblio::Thesaurus` thesauri files contain meta-information like the name of the base language, text encoding and rules about relations. See next subsection for more information.

2.2. Ontologies Completion

It is complex to maintain consistency in a big ontology. If we relate the term a with b, we always forget to edit b entry, and relate it with a.
`Biblio::Thesaurus`(Simões and Almeida, 2002) solves this problem with the ability to define mathematical properties about relations like the *inversion* property. If the user defines these properties, the module will be able to complete automatically the ontology.
If we use the following ontology defining INSTances as inverse of IOF (Instance of),

```
%inv INST IOF

feline
INST cat, tiger, panther
```

the tool will output the following completed ontology:

```
%inv INST IOF

feline
INST cat, tiger, panther

cat
IOF feline

tiger
IOF feline

panther
IOF feline
```

2.3. Merging Ontologies

One of the main tasks when building an ontology from different resources, namely, from thesauri, is to conciliate them.

Merging thesauri is trivial for terms which just appear in one of the thesaurus. When there is the same term in two thesauri, we need to check if we can conciliate them. This is the harder task. We all know that a term can have many senses. It is true that thesaurus keep different terms for different concepts. But two thesauri can have the same term for different concepts.

To solve this problem, our approach is the definition of heuristics to try find incompatible terms. For example, *if two terms are instances of the same concept, then they should mean the same*. While some of these heuristics can be hard-coded and used for any thesaurus, some need to be defined accordingly with the thesauri being joined.

In case of conflicts we added the concept of namespace. If the same term a is used for two different concepts, we take a "*instance of*" or an "*border term*" related term[3] (b), and use it as namespace, thus creating a concept represented by a and the other by $b :: a$. While this concept is important, we will not discuss it further in this article.

If there is no conflict we have to join the information. For standard relations it is a simple task: just calculate the union of related terms. In case of external relations we have to choose one of them.

For languages, we should keep the two (or more) translations. While this is planned, at the moment we choose only one of them. The problem with keeping more than one translation if the inversion of thesaurus as discussed on subsection 2.5..

2.4. Adding a new Language

Other problem on reusing thesauri available in the Internet is their lack of the Portuguese translations.

This means hard hand work adding those translations. We done that for the UNESCO thesaurus with help of some students. That resulted in a lot of hours of work, and in a very strange thesaurus as different students ended up translating concepts to the same identifier, or translating them in conflicting ways. The thesaurus is now being reviewed by a single person and we hope to have a good resource in the near future.

Meanwhile, we do not want to have all this work for every thesaurus we find in the Internet. So, we are developing a tool to translate thesauri in a semi-automatic fashion (the user can interact with the system, or just trust on its work).

This tool uses a set of external dictionaries to perform the translation. These dictionaries can be hand-made, or can be extracted automatically from parallel corpora(Simões and Almeida, 2003) or other multilingual thesauri we have which include the Portuguese language (as well as the now available UNESCO thesaurus).

The process of adding the new language can be followed by the change of the base language of the thesaurus (see subsection 2.5.). In any case, we obtain a new multilingual thesaurus from where we can extract monolingual thesauri structures.

[3] In this order, giving priority to the *instance of* relationship.

2.5. Changing an Ontology base language

The structure of a multilingual thesaurus assumes that there is a base language: the language which is available for all concepts in the thesaurus.

When constructing a resource for Portuguese, we want it to be the base language of our thesaurus. Because we want to use freely available thesauri from the web, we need a way to change their base language (which rarely is the Portuguese).

This process does not work when some entries of the thesaurus do not have a Portuguese translation. On these cases we are searching for the translation in a dictionary and, if we don't find it, marking the missing translations for future hand verification.

As an example, consider the following input with English as base language:

```
1  cat
2  PT gato
3  IOF feline

4  feline
5  INST cat
```

To use the tool we need to name the current base language (EN), to refer the new base language (PT) and the file to be processed.

```
```

The tool will return the new thesaurus with Portuguese as base language:

```
1  gato
2  EN cat
3  IOF [PT-EN:feline]

4  [PT-EN:feline]
5  INST gato
```

It is important to note that when we do not have the translation form a term, a special `[L1-L2:original]` term is created in order to keep the connections between concepts and to help in the process of finding doubts and missing translations (a set of functions is provided to extract this list of missing translations).

2.6. More complex example

In the following example, the table contains the following fields:

- name of a river;
- length;
- list of the countries;
- list of geographic places;

The template definition is straightforward:

```
1  %inv flows_thru traversed_by
2  %ext length

3  $1
4  iof rio
5  length $2
6  in $3
7  flows_thru  $4
8  %%

9  Danube:2860:Germany,...: Vienna, Budapest
```

Notes:

- line 1 – defines a new relation and its inverse
- line 2 – `length` co-domain is not a term

After running TabularThesaurus and making the thesaurus completion we get a ontology with this kind of entries:

```
1  Budapest
2  IOF city
3  IOF capital
4  DEF Budapest is capital-city of Hungary
5  IN Hungary
6  traversed_by Danube

7  Hungary
8  HAS Budapest
9  HAS Danube

10 city
11 INST Budapest, Lisbon, Vienna, Rome

12 ...
```

3. Recycling Taxonomies

Taxonomies are normally easier to find than Thesaurus or Ontologies as they are mathematically simpler. There is just one main relation (which in fact can be seen as the union of a set of relations) with an hierarchic relationship. Given these properties, it is possible to construct a tool to convert taxonomies into ontologies. Consider the following extract from a taxonomy:

```
1  Agronomy
2     Crop Science
3       Crop Production
4       Grain
5         Corn
6         Rice
7         Wheat
8       Tobacco
9     Horticulture
10      Floriculture
11      Turf or Ornamental Grass
12      Viticulture
13        Viniculture _USE_ Viticulture
14    Pedology _USE_ Soil Science
15    Pest Management
16      Forest Pest Management
17    Seed Production
```

We can easily transform this structure into the following ontology

```
1  Plant Sciences
2  NT        Plant Biology, Agronomy
3  NT      Plant Ecology
4  BT        IRIS_top_term

5  Agronomy
6  NT        Crop Science, Horticulture
7  NT      Seed Production
8  NT      Pest Management
9  BT        Plant Sciences, Agriculture

10 Grain
11 NT      Corn, Rice, Wheat
12 BT      Crop Science

13 Crop Science
14 NT        Soybeans, Grain, Tobacco
15 UF        Crop Production
16 BT        Agronomy

17 Pedology
18 USE     Soil Science
```

4. An Ontology builder: Tabular Thesaurus

It is easy to find list of terms in the Internet, like country names, capital cities, animals, and so on. These are normally found in tabular formats. For example, countries and their capital cities:

```
1  Lisbon:Portugal
2  Budapest:Hungary
3  Vienna:Austria
4  Madrid:Spain
```

This representation is very compact and very easy to edit. **TabularThesaurus** is a domain-specific language that uses as input a file with the table and a simple description of how each element should be transformed into a thesaurus entry. For instance,

```
1  $1
2  IOF city, capital
3  IN $2
4  DEF $1 is capital-city of $2

5  %%
6  Lisbon:Portugal
7  Vienna:Austria
8  Budapest:Hungary
```

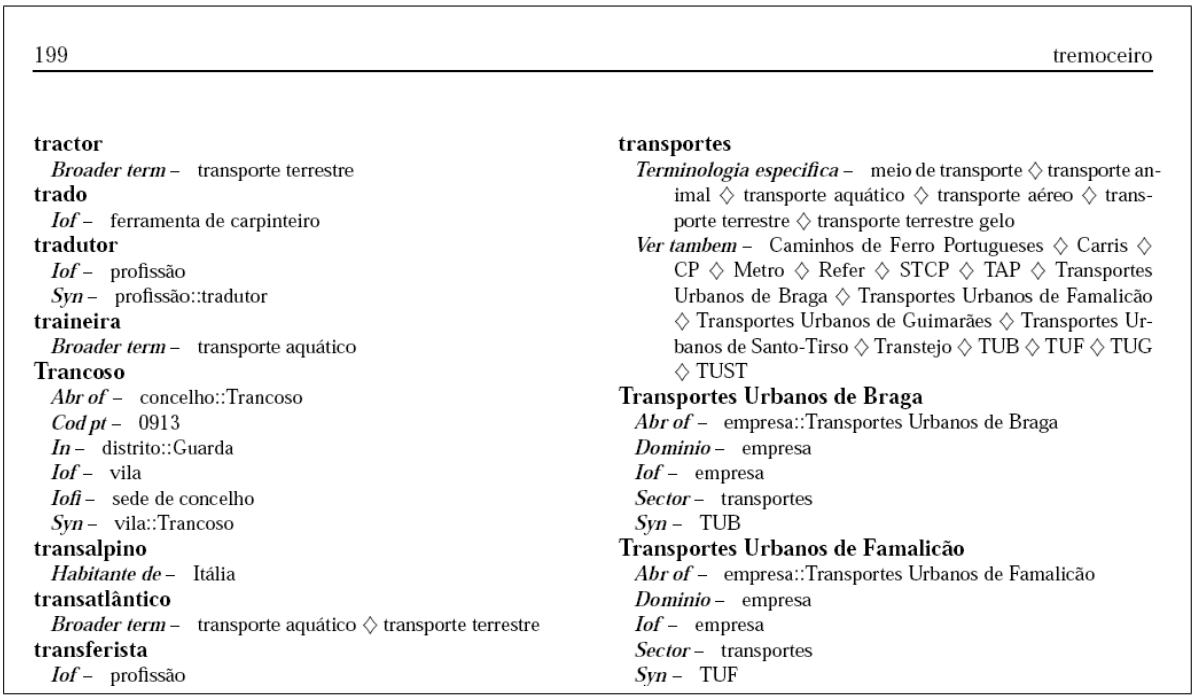
is a short way to define this thesaurus:

```
1  Lisbon
2  IOF city, capital
3  IN Portugal
4  DEF Lisbon is capital-city of Portugal

5  Vienna
6  IOF city, capital
7  IN Austria
8  DEF Vienna is capital-city of Austria
9  ...
```

5. Dictionary Views

Just to prepare a resource is not sufficient. The most important part in these kind of projects is to make the resources available in a suitable format, useful for the main user.



199 tremoceiro

tractor
Broader term – transporte terrestre
trado
Iof – ferramenta de carpinteiro
tradutor
Iof – profissão
Syn – profissão::tradutor
traineira
Broader term – transporte aquático
Trancoso
Abr of – concelho::Trancoso
Cod pt – 0913
In – distrito::Guarda
Iof – vila
Iofi – sede de concelho
Syn – vila::Trancoso
transalpino
Habitante de – Itália
transatlântico
Broader term – transporte aquático ♢ transporte terrestre
transferista
Iof – profissão
transportes
Terminologia especifica – meio de transporte ♢ transporte animal ♢ transporte aquático ♢ transporte aéreo ♢ transporte terrestre ♢ transporte terrestre gelo
Ver tambem – Caminhos de Ferro Portugueses ♢ Carris ♢ CP ♢ Metro ♢ Refer ♢ STCP ♢ TAP ♢ Transportes Urbanos de Braga ♢ Transportes Urbanos de Famalicão ♢ Transportes Urbanos de Guimarães ♢ Transportes Urbanos de Santo-Tirso ♢ Transtejo ♢ TUB ♢ TUF ♢ TUG ♢ TUST
Transportes Urbanos de Braga
Abr of – empresa::Transportes Urbanos de Braga
Dominio – empresa
Iof – empresa
Sector – transportes
Syn – TUB
Transportes Urbanos de Famalicão
Abr of – empresa::Transportes Urbanos de Famalicão
Dominio – empresa
Iof – empresa
Sector – transportes
Syn – TUF

Figure 2: Ontology LaTeX view

Is it not just important to publish, but also to do that early. To publish shows the utility of things earlier. If you need to increase the budget of your project, for instance, it will help if you can show how things will look at the end of the project. Also, people start looking at the published documents, giving feedback. This feedback will allow you to develop better resources, best views as well as to improve usability of the views made available. Finally, to publish help you to get friends and collaboration for you projects. Lot of projects began with little information and now include gigabytes or data. If you wait to have a good quantity of information before publishing it you will end up with a slower grow curve.
To make our ontologies available we are building:

- PDF views (through LaTeXand DiCTeX, as shown in image 2);
- Multi-file static HTML. This generation tool makes a HTML document for each term in the dictionary (see image 3);
- HTML dynamic sites;
- Bilingual-dictionaries (with domain information);
- StarDict, wiseDict, xdxf and DictD dictionary files;

Hungria

Francês: *Hongrie*
Espanhol: *Hungría*
Inglês: *Hungary*
Alemão: *Ungarn*
Broader term:
Europa de Leste
PECO
Países da CMAE
países Comecon
países NATO
países da OCDE
países do Conselho da Europa
países do Pacto de Varsóvia
Related term: Grupo de Visegrado
Codigo_tld: hu
Fullname: República da Hungria
Gentilico: húngaro
Has:
Budapeste
rio Danúbio
In: Europa

Figure 3: Ontology HTML view

6. Conclusions

The TabularThesaurus tool has proved to be an efficient way to define ontologies, and with a syntax easy to be used by normal users.
In our experiments, the use of a few word lists, thesauri and multilingual terminologies resulted in a dictionary with about 100 000 entries. The tool also creates the 100 000 HTML files needed to navigate on the ontology, as well as 600 pages on a PDF file for the printed dictionary.
In this experiment appeared some cases of ambiguity that we did not solve yet. As discussed before, the resolution of this problem needs the use of some heuristics we are yet dealing with.
The quality of the resulting dictionary highly depends on the quality of the resources used. Some of the resources

used introduce some noise, and should be previously filtered manually or semi-automatically.

6.1. Future Work

Rule-based forward chaining completion
In order to add some (very simple) inference mechanism we propose a Rule-based forward chaining approach.
Consider the following examples:
The capital of a country is a city:

$$A \text{ is\_capital } B \wedge B \text{ iof } Country \Rightarrow add(A \text{ iof } city)$$

The relation **iof** (instance of) is more specific than the relation textbtboarder term. So, if we have both relations we can delete boarder term:

$$A \text{ iof } B \wedge A \text{ bt } B \Rightarrow delete(A \text{ bt } B)$$

The implementation of this kind of rules in `Biblio::Thesaurus` should be easy, and is planned for real soon.

Relations ordering
Another special kind of rule is the relation ordering. If we define that some relations are more specific than others (for instance, *instance of* is more specific than *border term*) we can simplify the thesaurus removing one in favour to the more specific.

7. References

European Community. 1991. *European Education Thesaurus*. Luxembourg.

EuroVoc. 2004. Eurovoc v4.1 thesaurus. http://europa.eu.int/celex/eurovoc/.

ISO 2788. 2002. *Guidelines for the establishment and development of monolingual thesauri*. International Organization for Standardization.

ISO 5964. 2002. *Guidelines for the establishment and development of multilingual thesauri*. International Organization for Standardization.

Alberto Manuel Simões and José João Almeida. 2002. Library::⋆ — a toolkit for digital libraries. In *ElPub 2002 - Technology Interactions*.

Alberto M. Simões and J. João Almeida. 2003. Natools – a statistical word aligner workbench. *SEPLN*, Sep.

UNESCO, editor. 1995. *UNESCO Thesaurus: A Structured List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Social and Human Science, Culture, Communication and Information.* Unesco Publishing, Paris.