Compiling terminological data using comparable corpora: from term extraction to dictionaries

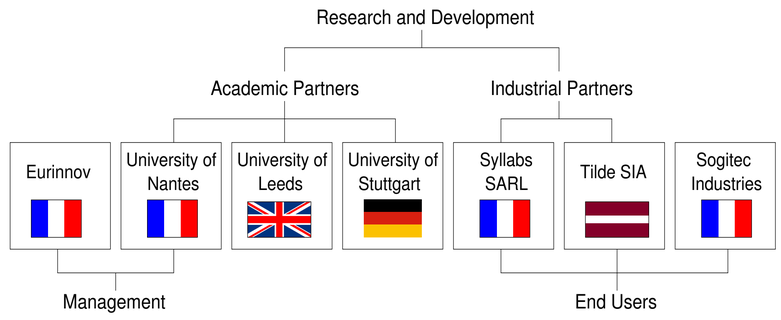
Marion Weller, Anita Gojun, Ulrich Heid
IMS - Universität Stuttgart
{wellermn|gojunaa|heid}@ims.uni-stuttgart.de

Béatrice Daille, Emmanuel Morin
LINA - Université de Nantes
{Beatrice.Daille|Emmanuel.Morin}@univ-nantes.fr

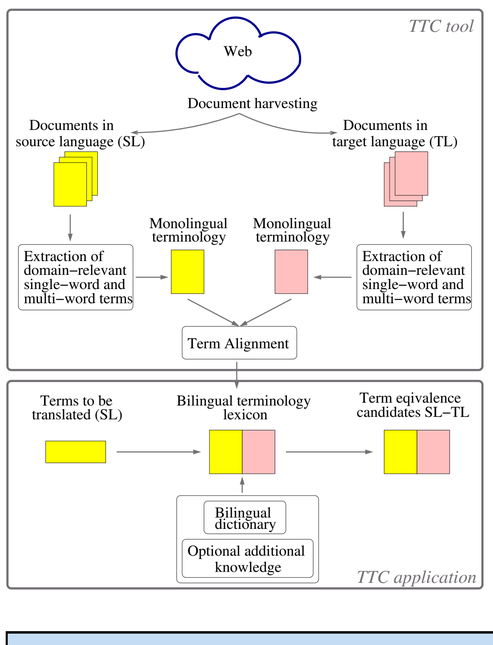
Overview

- Current situation: "Terminology Bottleneck" in translation
  - few resources for automatic bilingual terminology acquisition
  - lack of established terminologies for new/upcoming domains
- TTC solution: Term extraction from comparable corpora
  - semi-automatic tool chain
  - languages: DE, EN, ES, FR, LV, RU, ZH + selected language pairs
- Project Goal: Development of tools for corpus crawling, monolingual term extraction and bilingual term alignment
- Philosophy of tool development: Assessment of slim solutions: as little linguistic knowledge as possible

Project Partners:



Terminology Processing Chain



(1) Collecting domain-specific texts

- Thematic Web crawler for collecting documents from the Web [de Groc, 2011]

(2) Monolingual term extraction

- Input: monolingual crawled **domain-specific** texts
- Pre-processing: tokenizing – tagging – lemmatizing
- Monolingual extraction of **single-word** and **multi-word** term candidates
- Identification of **domain-relevant terms** using frequency-based and statistical approaches

(3) Identifying term variants

- Grouping **related terms** using pre-defined language-specific variation patterns
- **Output:** Groups of **synonymous and related monolingual term variants**

(4) Bilingual term alignment

- **Idea:** for a given source language term, find a translation in target term lists
- **Input:** bilingual general language dictionary + source and target language terms
- **Output:** bilingual domain-specific terminology
- Two approaches: **lexical strategy** and **context vector strategy**

Tools

- OpenSource UIMA-based application: **TTC Term Suite** code.google.com/p/ttc-project/
- Term extraction Web-service http://greenhouse.syllabs.com/ttc/

Term alignment: Lexical strategy

Method

- **Individually translate** the parts of a multi-word term
- **Combine** all translation possibilities
- **Compare** generated translation candidates with target language terms

Example

Input term	elektrisches$_{ADJ}$ Netz$_N$		
Lexicon look-up	elektrisch	→	électrique
	Netz	→	filet
	Netz	→	rets
	Netz	→	réseau
	Netz	→	secteur
Recombine & compare	filet électrique		not in target term list
	rets électrique		not in target term list
	réseau électrique		in target term list
	secteur électrique		in target term list
Output	elektrisches$_{ADJ}$ Netz$_N$	→	reśeau$_N$ électrique$_{ADJ}$ secteur$_N$ électrique$_{ADJ}$

Term alignment of compounds

Compounds

- A considerable amount of German domain-specific terms (N, ADJ) are compounds
- Usually not contained in general language dictionaries
- Compounds are often translated as multi-word terms

Method Use a **compound splitter** in order to obtain pseudo multi-word terms, then apply the lexical strategy for term alignment [Weller & Heid, 2012].

Examples

Kabellänge$_N$ → Kabel$_N$ Länge$_N$ → longueur$_N$ du$_P$ câble$_N$
Mastsockel$_N$ → Mast$_N$ Sockel$_N$ → base$_N$ du$_P$ mât$_N$
Tierart$_N$ → Tier$_N$ Art$_N$ → espèce$_N$ animale$_{ADJ}$

Problems

- Random matches with target language terms: *Leiter Platte* → #*board of directors*
- This method fails for non-compositional words: *Windschatten* (*lee position*)

Term alignment: Context vector strategy

Method

- **Lexical context analysis:** terms and their translations tend to appear in the same lexical contexts
- **Context vectors:** for each term, count occurrence frequencies of lexical units within a window of n words: this is done for source and target language terms
- **Translate** lexical units of the context vector of the source language using the bilingual dictionary
- **Compare** translation of source context vector with target language vectors (e.g. cosine measure)
- Terms with the **most similar context vectors** are likely to be translations

Problems

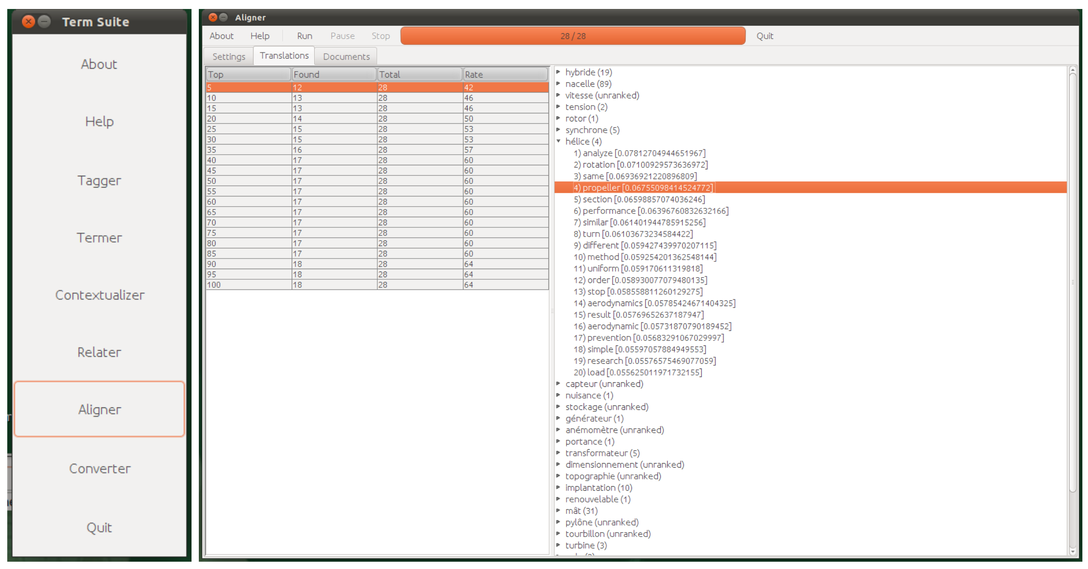
- Limited coverage of the general language dictionary
- This method is only suited for **single-word terms**

Ouput and evaluation

- For each input term, the top n results are shown, i.e. those target language terms with the most similar context vectors
- Results for aligning FR → EN:

	top 5	top 10	top 20
found alignments	42 %	46 %	50 %

Example: output of the alignment component in TTC Term Suite



References

[de Groc, 2011] Clément de Groc: Babouk: "Focused Web crawling for corpus compilation and automatic terminology extraction" in Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon, France, 2011.

[Weller & Heid, 2012] Marion Weller and Ulrich Heid: "Analyzing and Aligning German Compound Nouns" in Proceedings of LREC, Istanbul, Turkey, 2012.

Funding

The project has recieved funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement Number 248005

www.ttc-project.eu

scientific-contact@ttc-project.eu