Towards a Patents Translation System — Results and Perspectives

Ramona Enache Adam Slaski

MOLTO Project

May 3, 2011

Outline

The problem

- We are developing a hybrid translation system combining SMT and rule based strategies.
- The task is to translate patent claims from the biomedical domain.
- The main assumption is that the language of the claims is a limited subset of a natural language.
- Although, the language does not have to be easy. Consider the following example: *Mouse complement-fixing monoclonal antibody which (i) reacts with essentially all normal human peripheral T-cells but (ii) does not react with any of the normal human peripheral cells in the group comprising B cells, null cells and macrophages*.

Tokenizer

We have an on-purpose tokenizer—for treating compound noun phrases separated by hyphens, chemical compounds, etc.

Named-entity recognizer

For named entities recognition we use Stanford POS-tagger, which is state-of-the-art for general purpose POS-tagging and provides a very nice NE-recogniser, which proved to be useful for patents.
The work flow is first to identify named entities, then to substitute them in the main text with a dummy name (*AA* in this case) and store actual names in a separate file.

Number recognizer

We have also a special script for processing easily doubles, integers, percentages, ordinals, etc. As with the named entities we store actual values in a separate file.

Chemical compounds processing

For identifying chemical compounds we have a script that looks for words ending with certain prefixes (*ylic, rgic, inic*, etc.) or consisting of hyphens, brackets and commas (e.g. *7-[N-(1-carboethoxy-3-phenylpropyl)-(S)-alanyl-1,4-dithia-7-azaspiro[4.4]nonane-8(S)-carboxylic*). Also compounds consisting of more than one word are properly identified. As in previous cases compounds names are substituted with a dummy string and stored in a separate file.

Chemical compounds processing – part 2

Another problem is translating previously identified compounds. In most of the cases it is enough to split long names and translate them word by word. Although, *carbonic acid* should be translated to *acide carbonique*. Other changes in order are also possible and implementing them require more understanding of the structure of compounds.

Lexicon building

To extract the lexicon we use Genia POS-tagger which is profiled to work with texts written by scientists, specially form micro-biology domain. Genia does not provide named-entity recognition but has a very nice lemmatizer, which we use to obtain the base forms for the lexicon.
We use also the English Oxford Dictionary – abstract syntax and English concrete syntax - nouns, adjectives, adverbs and V2s.

Bootstrapping French lexicon

We bootstrapped French lexicon with the help of lexical tables from Moses and Google Translate.
Connected problems are:

- lack of lemmatisers for lexical tables,
- not reliable translation from Google Translate,
- possibly not lemmatized,
- corrected with French morphological dictionary Morphalou.

Grammar

For the grammar we extended the Resource Grammar with functions implementing constructions that occur in patent claims. Examples of such constructions are:

- *mouse antibody*;
- *complement-fixing antibody* (also spelled without hyphen);
- *hybridoma formed by a fusion*;
- *antibody according to claim 1*.

More about grammar

The grammar is suffering from a huge number of ambiguities. A lot of them was caused by the order of attaching objects to common nouns. E.g. *a big table in your office.* An attempt to reduce them was made by adding a hierarchy of common nouns. The idea is that on different levels different objects may attache to a common noun.

1. proper names (*class lgg, Lambda calculus*) and prepositional groups (*antibody in the group*);
2. adjectives;
3. relative clauses.

That reduces ambiguities in some cases even hundreds of times, but doesn't solve the problem in general.

Coverage

By now coverage oscillates about 15%.

Disambiguation

We plan to disambiguate parsing results at this point.

Restoring entities

Entities must be restored in a correct order.

Disambiguation grammar

There may be a lot of ambiguities. We plan to solve them in the following way:

- general rules—probabilities in the grammar,
- using the French aligned corpus for disambiguation,
- learn rules with Machine Learning.

Improving coverage of the grammar

We will

- write more rules,
- detect idioms (latin expressions, law jargon),
- detect prepositions and conjunctions which are specific to patents and extend the lexicon with them.

Dealing with large chemical compounds

SMT is lost dealing with compounds like *2,6-dimethyl-4-(3-nitrophenyl)-1,4-dihydropyridine-3,5-dicarboxylic acid-3-[1-(1-phenylethyl)-4-piperidinyl]-ester-5-methyl ester hydrochloride*.

Syntactical correctness

SMT seems not to preserve agreement in gender (present in French, not in English).

- Composition pharmaceutique selon la revendication 1 , *dans lequel* l' ibuprofène est (S) -ibuprophène .

Phonetic mutations

Because of lack of syntactic structure, SMT doesn't handle well phonetic mutations either.

- Utilisation *de le* 5-fluorouracile ...

Translation of Doubles

SMT doesn't handle the specific conventions for writing doubles:

- English - 1.5
- French - 1,5

Evaluation

We have a consultant on a chemistry department who evaluated the partial results yesterday.

The end

Thank you for your attention!