Simplification of Patent Claim Sentences for Their Paraphrasing and Summarization

Nadjet Bouayad-Agha,
Gerard Casamayor,
Gabriela Ferraro
Barcelona Media & Universitat Pompeu Fabra
Barcelona, Spain
{*firstname.lastname*}@upf.edu

Leo Wanner
ICREA & Universitat Pompeu Fabra
Barcelona,Spain
leo.wanner@icrea.es

Abstract

We present an approach to patent claim simplification which segments claim sentences into clausal discourse units, transforms them into complete sentences, establishes coreference relations and builds a discourse structure between discourse units. The four stages are necessary to allow for the syntactic analysis of otherwise unparsable claim sentences and their regeneration using discourse structure and coreference relations in order to ensure the production of a cohesive and coherent paraphrase/summary.

Motivation

In order to facilitate the comprehension of patent claims, which are very hard to read due to their complex linguistic style, we developed a rule-based paraphrasing and summarization module (Wanner et al. in press) that consists of three main components: the claim simplification component, the parsing component, and the regeneration component. In what follows, we focus on the claim simplification component. Claim simplification segments claim sentences into clausal discourse units, transforms them into complete sentences, establishes coreference relations and, finally, builds a discourse structure between discourse units. The four stages are necessary to allow for the syntactic analysis of otherwise unparsable claim sentences and their regeneration using discourse structure and coreference relations in order to ensure the production of a cohesive and coherent paraphrase/summary.

Simplification procedure

The simplification and discourse structure derivation procedure consists of the following sequence of processing stages:

1. *POS Tagging and chunking* using Schmid's (1994) TreeTagger with its off-the-shelf English parameters.

2. *Segmentation* of each claim using the best of a set of machine-learning (ML) and rule-based (RB) strategies based on a gold standard corpus of manually segmented claims. The gold standard consists of 1011 claims (6723 segments) from Optical Recording Device (OR) patents and 486 claims (3101 segments) from Machine Tool (MT) patents (the κ statistic of the agreement between the segment annotators over a subset of these claims was 79,8%). For the ML-experiments, we applied Weka's J48 decision tree learner (Witten and Frank 2005). To ensure an optimal tuning of the learner, we used a part of the gold standard, namely 581 claims (3942 segments) from the OR domain, as *development corpus* on which we ran J48 using 10 fold cross-validation. The remaining claims in both domains formed the *test corpus* (see the evaluation, next section). The vector of basic features comprised POS, chunks, single keyword and punctuation information on a window size of 9 words.

For RB-segmentation, we experimented with a variety of strategies on the development corpus. The best strategy used as information semi-colon, comma, and about twenty representative lexical markers and expressions typical of the patent genre such as *in which*, *characterized in that*, *so as to*, and other more ambiguous ones such as *and, for, by* when followed by a VP.

3. *Coreference Resolution*: Segmentation is followed by NP coreference resolution that relies upon the patent characteristic NP-repetition.[1]

4. *Clause Structuring*: During clause structuring, the segments are related in terms of subordination, coordination or juxtaposition to form a tree. The algorithm searches for the best clause structure in a space restricted by a set of weighted rules, each of which is further pondered by a set of constraints that ensure syntactic correctness and global coherence of the tree under construction. The rules encode the fundamental features for the identification of coordination, subordination and juxtaposition relations between spans. Each rule R is a quadruple $<F_1,F_2,W_0,W_c>$, where F_1 and F_2 are the features describing the left span S_1 and right span S_2 that are to be joined; W_0 is the rule's initial weight; and W_c is a set of weighted constraints that apply to the span features. As span features, we use, e.g., *punct*uation preceding the segment, *coord*ination, *subord*ination, and *synt*actic category (mainly Sentence, NP or VP) at the beginning of the segment. For instance, the rule for coordination of subordinations applies

[1] We obtained a performance of 79% for coreference resolution when evaluating the NP-repetition algorithm on 30 manually annotated coreference relations from both the OR and the MT domain.

between spans whose features are F_1={*coord*=-,*subord*=+} and F_2={*punct*=',',*coord*=+,*subord*=+}. The initial weight of the rules is currently set to 1, apart from a fall-back rule which is set to 0.001 as it must only apply if all others fail.[2] Based on a development corpus of 8 claims (104 segments), we have manually identified and adjusted 9 weighted constraints of the kind: 'spans start with the same syntagm: YES=1.0/NO = 0.5' (used to assess the likelihood of coordination).

To search amongst the various possible syntactic trees, we use a variation of a local beam search algorithm.

5. *Projection of the clause structure onto the discourse structure*: The coordinated constructs are first flattened in order to account for *n*-ary relations, then nodes of the tree are enriched by nucleus/satellite labels and discourse relations.

Evaluation of the main stages of the approach

In what follows, we report on the evaluation of the two main stages, segmentation and clause structuring. For segmentation, we used a strict evaluation that counts bijective 1:1 segment alignments. The results presented in Table 1 show that both ML and RB approaches perform quite well and are not far-off one another.

	p	*r*	*f*
Development corpus, OR:			
Rule	71%	63%	66.7%
J48	79%	68%	73%
Test corpus, MT:			
Rule	66%	60%	62.8%
J48	70%	64%	66.8%
Test corpus, OR:			
Rule	65%	59%	61.8%
J48	67%	62%	64.4%

Table 1: Evaluation results for segmentation

Our test corpus for clause structuring consisted of 14 claims from the OR domain (144 segments) and 15 claims from the MT domain (156 segments). We performed two evaluation runs, one using as input the raw claims, and the other the manually segmented and coreferenced claims. As baseline, we used right branching. For evaluation, we counted the number of identical spans between the automatic and manual structuring. In order to be able to compare spans from the raw input with spans from the gold standard, we automatically map each segment of the raw input to its corresponding gold standard segment. The results are shown in Table 2. The count of spans in the table does not include the top span, which would always be counted as correct. The best accuracy has been achieved with perfect input in the OR domain with an *f*-score of 61%; clause structuring in the same domain achieved 44% from raw input and 32% with the right-branching strategy as baseline.

[2]The fall-back rule ensures that complete trees are always constructed. Furthermore it is meant to take into account intra-clausal constructions such as appositions.

		p	*r*	*f*
Gold	MT	51%	51%	51%
	OR	62%	61%	61%
	Av.	56%	55%	55%
Raw	MT	49%	40%	44%
	OR	47%	41%	44%
	Av.	52%	36%	42%
Baseline	MT	38%	41%	39%
	OR	30%	35%	32%
	Av.	34%	38%	35%

Table 2: Evaluation results for clause structuring

Conclusions

We have developed an approach to the simplification of patent claims that avoids any loss of information by uncovering the claim's discourse structure and coreference relations for use in the subsequent regeneration stage. The results of the segmentation and clause structuring are promising.

So far, not many works target the problem of patent claim simplification and patent claim discourse structure derivation. Sheremetyeva (2003) discusses the problem of syntactic analysis of patent claims, but does not address the challenges of simplification and discourse structure analysis. Shinmori et al (2003) derive the discourse structure in Japanese patent claims. However, in Japanese patent claims, multiple sentences are coerced into one complex sentence which facilitates the use of a standard cue phrase based approach. This is not the case in English claims.

References

Schmid, H. 1994. Probabilistic part–of–speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 44–49.

Sheremetyeva, S. 2003. Natural language analysis of patent claims. In *Proceedings of the ACL Workshop On Patent Corpus Processing*.

Shinmori, A.; Okumura, M.; Marukawa, Y.; and Iwayama, M. 2003. Patent processing for readability. structure analysis and term explanation. In *Proceedings of the Workshop on Patent Corpus Processing held at the ACL Meeting*, 56–65.

Wanner, L.; Bott, S.; Bouayad-Agha, N.; Casamayor, G.; Ferraro, G.; Graën, J.; Joan, A.; Lareau, F.; Mille, S.; Rodríguez, V.; and Vidal, V. in press. Paraphrasing and multilingual summarization of patent claims. In Wanner, L.; Brügmann, S.; and Diallo, B., eds., *PATExpert: A Next Generation Patent Processing Service*. IOS Press, Amsterdam.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.