Multilingual Summarization in Practice: The Case of Patent Claims

Simon Mille[1] and Leo Wanner[1,2]

[1]Department of Information and Communication Technologies, Pompeu Fabra University,
Ocata, 1, 08003 Barcelona, Spain
[2]Catalan Institute for Research and Advanced Studies (ICREA),
Lluis Companys, 23, 08010 Barcelona, Spain
simon.mille@upf.edu, leo.wanner@icrea.es

Abstract. Hardly any other type of textual material is as difficult to read and comprehend as patents. Especially the claims in a patent reveal very complex syntactic constructions which are difficult to process even for native speakers, let alone for foreigners who do not master well the language in which the patent is written. Therefore, multilingual summarization is very attractive to practitioners in the field. We propose a multilingual summarizer that operates at the Deep-Syntactic Structures (DSyntSs) as introduced in the Meaning-Text Theory. Firstly, the original claims are linguistically simplified and analyzed down to DSyntSs. Then, syntactic and discursive summarization criteria are applied to the DSyntSs to remove summary irrelevant DSyntS-branches. The pruned DSyntS are transferred into DSyntSs of the language in which the summary is to be generated. For the generation of the summary from the transferred DSyntSs, we use the full fledged text generator MATE.

Keywords: patent claims, summary, machine translation, Meaning-Text Theory, Deep-Syntactic Structure.

1 Introduction

Hardly any other kind of text material is as notoriously difficult to read and comprehend as patents. This is first of all due to their abstract vocabulary and very complex syntactic constructions. Especially the claims in a patent are a challenge: in accordance with international patent writing regulations, each claim must be rendered in a single sentence. As a result, sentences containing more than 250 words are not uncommon; consider a still “rather short” claim from the patent EP0137272A2:

(1) An automatic focusing device comprising: an objective lens for focusing a light beam emitted by a light source on a track of an information recording medium; a beam splitter for separating a reflected light beam reflected by the information recording medium at a focal spot thereon and through the objective lens from the light beam emitted by the light source; an astigmatic optical system including an optical element capable of causing the astigmatic aberration of the separated reflected light beam; a light detector having a light receiving surface divided, except the central

> portion thereof, into a plurality of light receiving sections which are arranged symmetrically with respect to a first axis extending in parallel to the axial direction of the optical element and to a second axis extending perpendicularly to the first axis and adapted to receive the reflected beam transmitted through the optical element and to give a light reception output signal corresponding to the shape of the spot of the reflected light beam formed on the light receiving surface; a focal position detecting circuit capable of giving an output signal corresponding to the displacement of the objective lens from the focused position, on the basis of the output signal given by the light detector; and a lens driving circuit which drives the objective lens along the optical axis on the basis of the output signal given by the focal position detecting circuit.

A sentence of this length and complexity is difficult to process even for native speakers of English, let alone for foreigners who do not master English well. Given that professionals have to sift through the claims of a large number of patents returned as response to a search in a patent DB (which makes a quick assessment of the relevance of patent essential), it is not surprising that multilingual summarization of patent claims is very attractive to practitioners in the field. Nonetheless, only little work has been done so far in the area; cf. as an example [1], who proposes a reading aid based on the segmentation of claims into smaller and simpler sentences. The focus has been on the machine translation – especially in the light of the recently dramatically increased prominence of patents in languages not widely spoken in the West (e.g., Korean and Chinese).

As far as summarization of patent material is concerned, up to date, the overwhelming share of it is manual.[1] One explication for this unsatisfactory state of affairs is that the peculiarities of the genre of patent claims require new approaches to summarization: the application of *surface level criteria* such as term frequency, position etc., *term level criteria* such as similarity, word co-occurrence, etc. or *text* or *discourse level criteria* such as lexical chains, discourse relation trees, etc. to claims in their original form is not appropriate. The linguistic style of patent claims requires a novel summarization strategy that implies prior segmentation, simplification and text structure and discourse analysis.

We present an experimental rule-based module for the production of multilingual summaries from English patent claims developed in the framework of the PATExpert patent processing service.[2] The target languages are French, Spanish and German. The module currently undergoes an extensive evaluation and further extension. However, already in it present state it shows promising performance.

The remainder of the paper is structured as follows. In the next section, we assess the different ways to address summarization of patent material and briefly outline our approach to multilingual summarization. Section 3 presents the strategy in more detail. In Section 4, the evaluation of the performance of both summarization and multilinguality is presented. Section 5, finally, summarizes the main points of our work and gives hints to related work.

[1] Thomson Derwent is the world leading company in services for semi-manual patent abstracting; see http://scientific.thomson.com/derwent/

[2] PATExpert has been partially funded by the European Commission under the contract number FP6-028116. See [2] for a general presentation of the PATExpert service.

2 How to Do Multilingual Summarization of Patent Claims?

The abstract vocabulary of patent claims and their complex linguistic structures make a deep analysis needed for *abstraction* very hard, such that linguistically less challenging shallow summarization seems more promising.

One option is to exploit the claim tree structure, which defines the dependency between claims, cutting branches of the tree at depth *n* in accordance with the length of the summary desired by the user. This strategy reflects that claims at depth *n* are more general (and thus more relevant to the summary) than claims at depth *n+1*. But it does not increase the readability of the summary and is still very difficult to translate. Therefore, it is more appropriate to identify claim chunks (rather than entire claims) as relevant/irrelevant to the summary.

Since standard parsing algorithms are not able to cope with a reasonable outcome with sentences of such a length, a prior two-step simplification procedure of the original is needed: (i) segmentation into simpler chunks and (ii) repair of chunks which are not grammatical clauses by introducing missing constituents or referential links, or by modifying available constituents. The output of the simplification can serve for two *extraction* based summarization strategies: (a) discourse structure oriented summarization; (b) syntactic structure oriented summarization.

Discourse structure oriented summarization as proposed by [3] uses the depth of the subtree "controlled" by an element of a discourse relation in the sense of the Rhetorical Structure Theory [4] – under the assumption that the nucleus of a relation controls an elementary tree formed by the nucleus and satellite of a relation. See [5] for the application of this strategy to the summarization of patent claims.

The syntactic structure based summarization often uses syntactic dependency criteria which indicate the importance of syntactic tree branches, drawing on dependency relations [6,7]. To the best of our knowledge, the syntax oriented strategy has not been applied so far to patents.

In PATExpert, three different summarization strategies are implemented: (i) a strategy based on the claim structure, (ii) a strategy based on the discourse structure, and (iii) a strategy based on the deep-syntactic (or shallow semantic) structure. Cf. Figure 1 for the architecture of the summarization module.

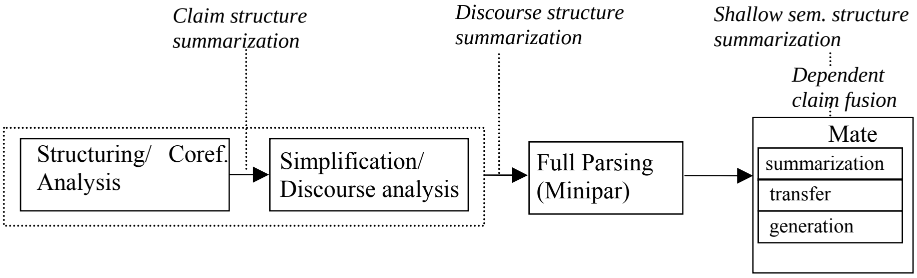

Claim structure summarization
Discourse structure summarization
Shallow sem. structure summarization
Dependent claim fusion
Structuring/ Coref. Analysis
Simplification/ Discourse analysis
Full Parsing (Minipar)
Mate
summarization
transfer
generation

Fig. 1. Architecture of the multilingual summarization module

The shallow semantic (or deep-syntactic) structure summarization is most suitable for multilingual summarization. It presupposes two preprocessing stages: (a) claim dependency structure determination, simplification, and discourse analysis, and (b) full parsing of the simplified claim sentences. For parsing, we use MINIPAR [8]. Despite some shortcomings such as systematic right-attachment, we chose MINIPAR since it produces syntactic structures which roughly correspond to the Surface Syntactic Structures (SSyntSs) of the linguistic framework underlying the linguistic workbench MATE [9] we use for generation: the Meaning-Text Theory, MTT [10].

The summarization and multilingual transfer stages are performed on the Deep-Syntactic Structures (DSyntSs) of the MTT, such that prior to these stages, the MINIPAR structures are mapped onto SSyntSs and the SSyntSs onto DSyntSs; for details on the preprocessing stages, see [11]. The abstract nature of the DSyntS, which eliminates the surface-syntactic idiosyncrasies of the linguistic constructions, ensures, on the one hand, quasi-semantic criteria for summarization, and, on the other hand, simplified transfer between the structures of different languages; cf., e.g., [12].

3 Multilingual Summarization of Patent Claims

Starting from the DSyntSs of the simplified claims, the multilingual summarization of patents consists of the following steps: (1) summarization of the original claims, (2) transfer of DSyntS of the source language to the target language, (3) generation of the summary in the target language.

3.1 Deep-Syntactic Summarization

The summarization criteria are based on specific patterns recognized within the input DSyntS. These patterns trigger the application of summarization rules from the summarization grammar defined in MATE. Consider some of these patterns and the effect of the application of the corresponding summarization rules, namely removing of the chunks (in reality, branches of the DSyntS) that appear in brackets:

1. A noun has a postponed attribute:
   (a) *The optical component is a shading member* [***arranged*** *near the optical axis around the aperture plane of the optical system*].
   (b) *The recesses are formed in the upper face and extend from a land surface* [***adjacent*** *to said cutting edge*].
2. A definite noun is modified by a full statement:
   (a) *An automatic focusing apparatus comprises* ***the*** *actuator* [***which*** *controls the focusing means depending upon the output of the phase detector*].
3. A noun in a dependent claim is modified by a “has-part” relation (in an independent claim, it can bear important information):
   (a) *A unitary ridge is formed on the top face* [***having*** *side surfaces constituting the first and second side chip deflector surfaces*].
4. A noun in a dependent claim is modified by a PURPOSE relation (*for* + Gerund):
   (a) *The apparatus comprises a lens* [***for*** *convert****ing*** *the light from the signal plane*].

5. A number appears in a sentence of a dependent claim:
   (a) *The reflective component-containing layer has a film thickness of 0.01µm to 0.5 µm.*[3]
   (b) [*The film thickness is* ***0.01µm to 0.09 µm***].

Once the DSyntSs are cleared of redundant information in the summarization stage, they are aggregated in that coordination conjunctions, ellipses, and relative clauses are introduced to produce a more natural, fluent text; for details, see [11].

3.2 Multilingual Transfer

Aggregated structures serve as a starting point for the multilingual transfer. The prior simplification guarantees that the source language DSyntSs are rather simple – with the effect that the mismatches between the source and target DSyntSs are minimized (for handling of the mismatches at the DSyntS-level of transfer, see [12]). As a result, the transfer becomes to a large extent a *lexical transfer*. The transfer procedure proper is preceded by word disambiguation.

Disambiguation. The disambiguation of words must be addressed in order to obtain the correct translation from the transfer dictionary (see below). For instance, the English OPEN can be translated by two French verbs S'OUVRIR and OUVRIR. Which one is correct depends on the number of semantic actants (one or two) of OPEN. In other words, S'OUVRIR and OUVRIR correspond to two different senses of OPEN: OPEN1 and OPEN2.

An important criterion for the disambiguation is the subcategorization information available in the dictionary. Several simple rules retrieve from the dictionary the right entry for the verb according to the number of syntactic actants found in the DSyntS:

(4) ?X {–I→?Y} | ¬?X–II→?N ∧ ¬?X.voice=passive ∧ disambiguation::(?X.dlex).(I).(lex)
⇒
?X {dlex=disambiguation::(?X.dlex).(I).(lex)}

The above rule states that if a node bound to the variable ?X has a DSynt actancial relation "I" with the node ?Y, but no relation "II"; if it is not in the passive and has an entry in the "disambiguation" dictionary, then the name ("dlex") of ?X is the value of the attribute "lex", which is the non-atomic value of the attribute "I" in the entry of "?X" in the disambiguation dictionary. Applied to OPEN this rule gives us OPEN1:

```
(5) open {dpos=V
          I = {lex=open_1
                    gp = {I = {dpos=N}}}
          II = {lex=open_2
                    gp = {I = {dpos=N}
                          II = {dpos=N}}}}
```

[3] In (5a), the first sentence is an independent claim which has a dependent claim that contains the second sentence.

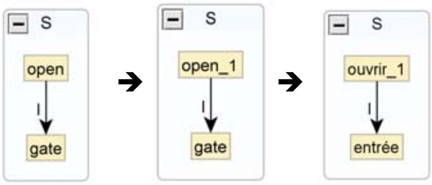
('gp' stands for "government pattern", i.e., valency structure). Monolingual dictionaries of this kind are available for each of the source and target languages (in our case: English, French, German and Spanish).

Multilingual Transfer Proper. The entries in the transfer dictionary have the following form:

(6) open_1 {V ={ FRE = {trad=ouvrir_1}
GER = {trad=öffnen_1}
SPA = {trad=abrir_1}}}

The transfer itself is simple and straightforward: the nodes of the disambiguated, summarized and aggregated DSyntSs are mapped almost one-to-one to the target DSyntSs by getting the translations from the transfer dictionary. Consider a very simple example sequence $\text{DSynt}_{\text{Eng}}$ ⇒Disambiguated $\text{DSynt}_{\text{Eng}}$ ⇒ DSynt_{Fr}:

(7)



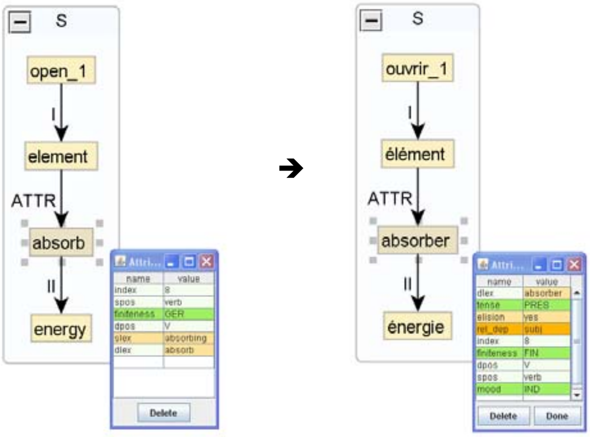
Most transfer rules are language-independent, but some of them preprocess the tree for the language-dependent surface-syntactic structural mismatches. For instance, the English construction N_1-V_{ing}-N_2[4] as in *signal processing circuit* is more naturally rendered in French or in Spanish via a relative clause pattern N_2-that-V-N_1. For DSyntS, this only means adding an attribute to the node of the verb which will trigger the introduction of the relative clause in SSyntS: relative pronouns are considered as a possible surface-syntactic manifestation of the ATTR DSynt-relation. The following rule establishes this equivalence:

(8) $?N_2$ { –?r→ ?X} | ?X.finiteness=GERUND ∧ ¬$?N_2$.dpos=Prep
⇒
?X {rel_dep=subj ∧ finiteness=fin ∧ tense=Pres ∧ mood =IND}

The value of the attribute "rel_dep" stands for the dependency relation that the relative pronoun has with its verbal governor; it indicates at the same time the presence of the relative pronoun in the SSyntS. This configuration is exemplified in (9) for *An* [*energy absorbing*] *element opens* vs. *Un élément* [*qui absorbe l'énergie*] *s'ouvre* 'An element which absorbs the energy'. The actual structural difference between the English and the French sentences is only surface-syntactic – such that it will appear only in the SSyntS, as shown in the next subsection.

[4] N_2 is the syntactical governor of the group, hence it is the top node in the rule below

(9)



name	value
index	8
spos	verb
finiteness	GER
dpos	V
slex	absorbing
dlex	absorb

name	value
dlex	absorber
tense	PRES
elision	yes
rel_dep	subj
index	8
finiteness	FIN
dpos	V
spos	verb
mood	IND

Multilingual Generation. During generation of the target language summary, the DSynt-SSynt transition is central. The DSynt-SSynt rules call the monolingual dictionaries in order to retrieve language-specific information such as governed prepositions, auxiliaries, pronominal status, etc. For instance, FROM is a preposition requested by the third actant of the English verb PREVENT. This preposition does not appear in the DSyntS and has to be generated in the SSyntS. Therefore, the corresponding preposition – if there is one – of the French equivalent EMPÊCHER must appear in the GP of EMPÊCHER in the monolingual dictionary; cf. (10). Similarly, thanks to the monolingual information, we know that OUVRIR1 is a pronominal verb:

```
(10) empêcher : verb_dt { //eng=keep/prevent
          elision = yes
          gp = {III = {dpos = V
                      rel = obl_obj
                      Prep = de}}}
```

```
ouvrir_1: verb { //eng:open_1
  lemma = ouvrir
  elision = yes
  pronominal =yes
  past_aux = être}
```

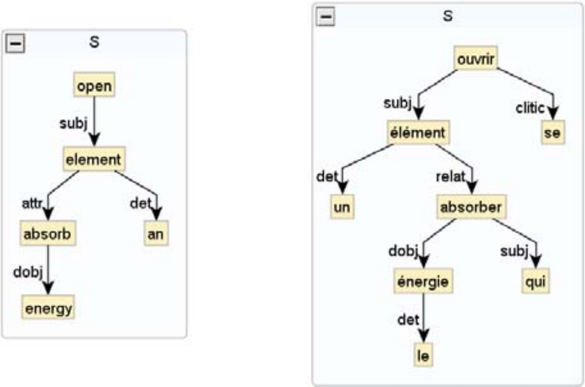
In French, the feature “pronominal” is realized by the clitic SE, which introduced by the following rule:

(11) ?V | lexicon::(?V.dlex).pronominal=yes ∧ language=FR
⇒
?V {–clitic→ ?X}

(11) checks the attribute “pronominal” in the entry for the verb in the lexicon. If the value is “yes”, a node “?X” and an edge “clitic” connecting ?X to the verb are created. The same kind of mechanism operates, for instance, for the introduction of relative pronouns and determiners.

(12) shows the SSyntSs that correspond to the DSyntSs in (9); the structural difference between English and French is now visible (the value ?r of the rel_dep attribute is “subj”).

(12)



The rest of the surface generation, i.e., the linearization and morphological processing of the lexical units is detailed in [11].

4 Evaluation of the Multilingual Summarization of Patent Claims

Given that no reliable unique evaluation metrics exists as yet for multilingual summarization, we performed a preliminary evaluation of our strategy of multilingual summarization from the perspective of the quality of the summary and from the perspective of the quality of the multilinguality.

The evaluation of the quality of our summary has been performed using ROUGE [13]. As baseline, we used the MS Word automatic summarizer (MSAS), with the summarization parameter set to 50%.

Out of a list of 50 patents that underwent simplification, 30 were randomly selected and summarized with our summarization module and MSAS. The summaries used as reference have been done by a patent specialist. Our summarization obtained an overall f-score of 61% over quadrigrams and trigrams, while MSAS reached 43%.

That we did not surpass 61% can be partially explained by the object/method dichotomy in some patent claims, which we cannot identify reliably in an automatic way. If a patent claim section contains claims referring to both the invented object and the method of applying this object, both kinds of claims tend to contain largely the same information. Human created reference summaries avoid the repetition of this information, while our module is currently not able to differentiate an object-related claim from a method-related one. Furthermore, it is worth noting that the evaluation that has been carried out so far does not take into account the quality of the text, for which a qualitative evaluation would be necessary.

For the evaluation of the quality of the multilinguality, we chose human evaluation in order to balance the purely statistical metrics of the ROUGE evaluation and to obtain some objective opinions from native speakers and experts. For this purpose, six native speakers were asked to rate twelve different claim descriptions in their native tongue produced by PATExpert with summarization switched off (such that only simplification, transfer and regeneration were effective) against the online-Google translation of the original claims as baseline.[5] Given that the recall of our multilingual generator is still very much hampered by the shortage of multilingual resources, we consider this evaluation a general indication of the potential of "deep" translation techniques when combined with the preprocessing of the claims.

The evaluation was based on a questionnaire which has been largely inspired by [14]. It consists of three categories: "intelligibility", "simplicity" and "accuracy". The first two deal with the quality of the transferred text; both have a five value scale. The third category, which has a seven value scale, captures how the content from the English input is conveyed in the transferred text. Due to the lack of space, we do not cite here the questionnaire itself. Table 1 shows the accuracy regarding each of the three quality categories for PATExpert and the baseline.

Table 1. Accuracy of the PATExpert Multilingual Summarizer against a baseline

	Google Translator (baseline)	PATExpert Multilingual Summarizer
Intelligibility	0,49	0,58
Simplicity	0,49	0,74
Accuracy	0,47	0,51

As expected, the complexity of our multilingual summarization module is much lower, hence the intelligibility is about 9% higher. But surprisingly, there is no significant difference regarding the accuracy of the two translations, which might show that no meaningful information is lost during the simplification stage compared to a non-simplified output.

5 Summary

From the practitioners' side, there is a high demand for multilingual summarization of patent claims. However, traditional approaches to summarization do not show the required performance due to the particular linguistic style and abstract vocabulary of the claims. In this paper, we proposed a strategy that makes use of a number of preprocessing stages for a prior linguistic simplification of the material and that integrates *de facto* the summarization into generation. This allows us, on the one hand, to perform the summarization at a rather abstract level and thus to use "deep" summarization criteria, and, on the other hand, to reduce the transfer to a large extent to lexical transfer. The results are encouraging. Still, the three central components involved in the process: summarization, transfer and generation, are continuously being extended and improved, such that in the full paper, we will be able to present evaluation figures that are likely to be considerably superior to those presented above. There are some related works. The most similar ones are the MUSI-summarizer by [15] and the summarizer within VERBMOBIL described in [16]. As our strategy,

[5] Since our goal was to evaluate the multilingual output of our system with the original claims as input, we consider it correct to run the Google translator on the original claims.

MUSI implies a deep analysis stage and a regeneration stage. However, MUSI's summarization strategy consists in sentence extraction using surface-oriented criteria (cue phases and positions of sentences). The analysis is applied to the extracted sentences and the resulting syntactic structures are mapped onto conceptual representations from which then the (possibly multilingual) summary is generated. [16] describes multilingual summary generation in a speech-to-speech system. The difference between their system and ours again mainly consists in the summarization strategy, with ours being considerably deeper.

Acknowledgments. Many thanks to the members of the TALN group at UPF as well as to all colleagues of the PATExpert Consortium for their valuable support.

References

1. Sheremetyeva, S. Natural language analysis of patent claims. In: ACL Workshop on Patent Corpus Processing, pp. 66 – 73. Sapporo (2003)
2. Wanner, L. et al. Towards Content-Oriented Patent Processing. World Patent Information. 30, 21–33 (2008)
3. Marcu, D. From discourse structures to text summaries. In: ACL/EACL Workshop on Intelligent Scalable Text Summarization, pp. 82–88. Madrid (1997)
4. Mann, W.C., Thompson, S.A. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, 8, 243–281 (1988)
5. Shinmori, A., Okumura, M., Marukawa, Y., Iwayama, M. Patent Processing for Readability. Structure analysis and Term Explanation. In: ACL Workshop on Patent Corpus Processing, pp. 56–65. Sapporo (2003)
6. Farzindar, A., Lapalme, G., Desclés, J-P. Résumé de textes juridiques par identification de leur structure thématique. Traitement automatique des langues. 45, 39–64 (2004)
7. da Cunha, I., Wanner, L., Cabré, T. Summarization of Special Discourse: The Case of medical articles in Spanish. Terminology. 13, 249–286 (2007)
8. Lin, D. Dependency-based Evaluation of MINIPAR. In: Workshop on the Evaluation of Parsing Systems, pp. 234–241. Granada (1998)
9. Bohnet, B., Langjahr, A., Wanner, L. A development environment for an MTT-based sentence generator. In: INLG 2000, pp. 260–263. Mitzpe Ramon (2000)
10. Mel'čuk, I. Dependency Syntax. SUNY Press, Albany (1988)
11. Mille, S., Wanner, L. Making Text Resources Accessible to the Reader: The Case of Patent Claims. In: LREC'08. Marrakech (2008)
12. Mel'cuk, I., Wanner, L. Syntactic Mismatches in Machine Translation. Machine Translation. 20, 81–138 (2006)
13. Lin, C. ROUGE: A Package for Automatic Evaluation of Summaries. In: ACL Workshop Text Summarization Branches Out, pp. 25–26. Barcelona (2004)
14. Nagao, M., Tsujii, J., Nakamura, J. The Japanese Government Project for Machine Translation. Computational Linguistics. 11, 91–109 (1985)
15. Lenci, A. et al. Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project. In: LREC'02, pp.1464–1471. Las Palmas (2002)
16. Alexandersson, J., Poller, P., Kipp, M., Engel, R. Multilingual Summary Generation in a Speech-To-Speech Translation System for Multilingual Dialogues. In: INLG-2000, pp. 148 –155. Mitzpe Ramon (2000)