Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus

Katharina Wäschle and Stefan Riezler

Department of Computational Linguistics, Heidelberg University
{waeschle, riezler}@cl.uni-heidelberg.de

Abstract. Statistical machine translation of patents requires large amounts of sentence-parallel data. Translations of patent text often exist for parts of the patent document, namely title, abstract and claims. However, there are no direct translations of the largest part of the document, the description or background of the invention. We document a twofold approach for extracting parallel data from all patent document sections from a large multilingual patent corpus. Since language and style differ depending on document section (title, abstract, description, claims) and patent topic (according to the International Patent Classification), we sort the processed data into subdomains in order to enable its use in domain-oriented translation, e.g. when applying multi-task learning. We investigate several similarity metrics and apply them to the domains of patent topic and patent document sections. Product of our research is a corpus of 23 million parallel German-English sentences extracted from the MAREC patent corpus and a descriptive analysis of its subdomains.

1 Introduction

Statistical machine translation (SMT) requires large amounts of parallel data on the sentence level to train translation and language models of high coverage. Best results are obtained if parallel data are available for the specific domain in question. Patent translation is particularily dependent on the availability of large in-domain parallel data sets for several reasons: Patent translation deals with documents that exhibit a highly specialized vocabulary, consisting of technical terms specific to the field of invention the patent relates to and legal jargon ("patentese") that is not found in everyday language. To maximize their coverage, patents are often intentionally vague and ambiguous. Furthermore, patents exhibit a complex textual structure of differently designated text fields. Some patent documents contain translations; however, in most cases only parts of the patent, namely titles and abstracts or claims, are multilingual, while for the largest document section, the description, no direct translation is available. This poses a potential data sparsity problem for patent translation.

In this paper we investigate possibilities for building resources for patent translation by extracting large amounts of parallel data from a multilingual patent document corpus, MAREC, and preparing the data for domain-oriented translation. Multi-task learning, for example, aims to enhance machine learning

performance by learning tasks on several separate but similar domains at the same time. Patents differ with respect to vocabulary – e.g. patents assigned to IPC section C (chemistry) tend to contain a large amount of chemical formula – and style – the patent title consists of a single noun phrase, while claims exhibit a complex phrasal structure. This induces different subdomains that can be viewed as tasks in multi-task translation. First experiments on multi-task patent translation on tasks defined by patent topic and document sections have been presented by [1].

The focus of this paper lies on the corpus construction and description, so we employ patent translation as a tool for similarity analysis. We train separate translation models on every subdomain and evaluate across domains in order to investigate similarities and differences between domains. Furthermore, we apply several information-theoretic similarity metrics to the topic dimension of IPC patent classification. While in general, every subdomain is best translated with a model trained on the respective domain, we find a correlation between similarity of IPC domains as measured by information-theoretic metrics and BLEU evaluation in cross-domain translation over IPC domains. This shows that information-theoretic measures can be used to select appropriate patent texts from related domains for augmenting training data. Furthermore, we show that due to particularities of the patent data language-specific preprocessing, such as German compound splitting, can be a key technique for German-English patent translation, especially for the translation of titles.

Product of our research is a corpus of over 23 million parallel German-English sentences from all IPC domains and text sections, sorted accordingly. Together with the descriptive analysis given in this paper, this forms an enabling resource for research on patent translation and tasks that build on translation, such as cross-lingual patent retrieval.

2 Related Work

Patent translation is an active research area that is fueled by benchmark testing workshops such as CLEF[1] and NTCIR[2]. NTCIR features a patent translation task for Japanese-English and Chinese-English patent documents. The data for the former task contains about 2 million sentence pairs that were automatically extracted from the description sections. The extraction method is described in [2]. It involves a pipeline architecture where in a first step length-based alignment scores ([3]) are used to propose sentence alignment candidates, which are then filtered using dictionary-based word translation scores. [2] also report results on patent translation experiments across IPC sections, showing that MT results are best when training and test sections coincide. Furthermore, pooling data from all sections for a maximum amount of training data achieved best results overall. [4] extract 160,000 Chinese-English sentence pairs using a pipeline of candidate

[1] Cross-Language Experiment Forum, http://www.clef-campaign.org

[2] National Institute for Informatics Test Collection for IR Systems, `http://research.nii.ac.jp/ntcir/`

sentence alignments that are filtered based on lexical translation scores. They do not report cross-section MT experiments.

The MAREC dataset is a superset of the patent retrieval data used in the CLEF-IP tracks, and has been deployed for stand-alone patent translation by [5], [6], and [7]. Again, an extraction procedure combining a candidate sentence alignment with a word-based translation filter is used. For example, [6] combines Gargantua ([8]) for sentence alignment with GIZA++ ([9]) for word alignment. Cross-section MT experiments for IPC domains are reported in [5] and [7], where the maximum-sized pool of combined data from all sections yields the best results, especially for language modeling. The sentence-parallel data extracted from MAREC for the experiments in [5], [6], and [7] is not publicly available. Furthermore, only data from abstracts and claims sections was extracted.

3 Structural and Topical Dimensions in Patent Text

We analyze patents with respect to the domain dimensions of both topic – the technical field covered by the patent – and structure – a patent's text sections – with respect to their influence on machine translation performance.

The topic dimension of patents is given by the International Patent Classification (IPC)[3] which categorizes patents hierarchically into 8 sections, 120 classes, 600 subclasses, down to 70,000 subgroups at the leaf level. Table 1 shows the 8 top level sections. A patent can be assigned to multiple IPC classes.

Table 1. IPC top level sections.

A Human Necessities
B Performing Operations, Transporting
C Chemistry, Metallurgy
D Textiles, Paper
E Fixed Constructions
F Mechanical Engineering, Lighting, Heating, Weapons
G Physics
H Electricity

In addition to the patent classification, we argue that patents can be subcategorized along the dimension of textual structure. Exemplary, the European Patent Convention (EPC) lays out the structure of a patent[4] in Article 78.1:

"A European patent application shall contain:
(a) a **request for the grant** of a European patent;
(b) a **description** of the invention;

[3] `http://www.wipo.int/classifications/ipc/en/`
[4] Highlights by the authors.

(c) one or more **claims**;
(d) any drawings referred to in the description or the claims;
(e) an **abstract**,
and satisfy the requirements laid down in the Implementing Regulations."

The textual elements of a patent are the title, which is specified in the request for grant, description, claims, and abstract. Examples for each text type can be found in table 2. The title is a short descriptive noun phrase, while the claim exhibits a particular sentence structure.

Table 2. Sample sentences from patent text sections.

title	Contact lense forming machine
abstract	Parameters for mold materials and important dimensions are also disclosed.
description	FIGS. 7 and 8 illustrate the final curvatures of a finished plus and minus lense, respectively.
claim	The machine as set forth in claim 2, wherein said lense holding element is secured to the fixed support and the cooperating tool element is mounted on the first pivotal support means.

4 Extraction of Parallel Text

Our work is based on the MAREC[5] patent data corpus. It contains over 19 million patent applications and granted patents from four patent organizations (European Patent Office (EP), World Intellectual Property Organization (WO), United States Patent and Trademark Office (US), Japan Patent Office (JP)), from 1976 to 2008 in a standardized format. We extract data for our experiments from the EP and WO subcorpora which contain multilingual patent documents that feature partial translations of the patent text between German, English and French (the EPO's official languages). We assume translated titles to be sentence-aligned by default, and define multilingual document sections, which are of similar length in both languages as parallel[6]. To extract parallel text sections, we first determine the longest instance of the respective section, if different document kinds[7] exist for a patent. Overall, we extracted 2,204,384

[5] `http://www.ir-facility.org/prototypes/marec`

[6] We compute the number of German tokens relative to the number of English tokens and keep parallel sections with a ratio larger than 0.7.

[7] A patent kind code indicates the document stage in the filing process, e.g., A for applications and B for granted patents, with publication levels from 1-9. See `http://www.wipo.int/standards/en/part_03.html`.

parallel titles, 291,716 parallel abstracts, and 735,667 parallel claims sections for the German-English language pair. However, there are no parallel descriptions.

The lack of directly translated descriptions poses a serious limitation for patent translation, since this section constitutes the largest part of the document. It has been shown that it is possible to obtain comparable descriptions from related patents that have been filed in different countries and are connected through the patent family id. [2] introduce this method to collect Japanese-English patent translations, [4] apply the same technique to Chinese and English patents. We transfer this approach to German-English patents and search the US collection of MAREC for documents sharing a family id with EP patents that feature a German description. We extracted 172,472 patents that were both filed with the USPTO and the EPO and contain an English and a German description, respectively. However, data extracted in this way is presumably less parallel than the directly translated sections. This is due to the different origination process: translations of document sections are drawn up during the application process and aim to stay close to the original text. The filing of an application for the same or a closely related invention in a different country might occur several years after the first publication, during which the invention might have been subject to changes. Due to differing regulations and application procedures there might be further modifications, amendments and omissions compared to the original document. [2] and [4] use a scoring function to determine good translations after performing sentence alignment, deploying an aggregate score of sentence length ratio, IBM Model 1 word translation probabilites and a lexical score using a dictionary. Since we do not have access to a domain-specific dictionary[8], we enforce the first two criteria for the selection of parallel sentences from descriptions in the sentence alignment process.

For alignment, we used Gargantua[9] ([8]), an unsupervised, language pair independent open source aligner, which implements the idea presented in [11]: a combination of both sentence-length-based alignment and a lexical translation model in a two-pass approach. The best alignment is calculated based on sentence length in the first pass, and recalculated with lexical word translation probabilities from an IBM Model 1 estimated on the preliminary alignments in the second pass. The algorithm is robust and can deal with asymmetrical translations, generating one-to-many/many-to-one and 1-to-0/0-to-1 alignments.

Before running the sentence aligner, we cleaned the extracted data and removed several forms of noise. Among the problems we observed were misspellings, hyphens left from line breaks, e.g. Wasser-\nverteilungsnetz, formatting tags such as <IMAGE> or <SEP>, and multiple or missing whitespace.

We eliminated formatting-related problems, such as multiple whitespace, line breaks, tags and hyphens. We did not attempt to correct misspellings, assuming that they will not hurt translation performance significantly. Another problem we observed with MAREC documents are mislabeled language attributes, e.g. a

[8] A patent domain dictionary might be created from high-quality multilingual data, for example titles.

[9] `http://gargantua.sourceforge.net/`

French abstract with English language label EN. We did not perform language detection and therefore cannot quantify the mislabelings, but they appear to be very infrequent; we rely on the sentence aligner, which uses lexical word probabilities, to filter out sentences in the mistaken language.

Sentence aligning requires further preprocessing of the input text, namely splitting the text into sentences and tokenization. We use language-specific, heuristic-based tools distributed with Europarl[10] ([10]), which resolve punctuation ambiguities[11] by using a list of known abbreviations and heuristics for each language, which we extended to include patent-specific abbreviations.

Table 3. Alignment statistics.

		de		en	
	output	input	percentage	input	percentage
abstract	720,571	780,161	92.36%	938,117	76.81%
claims	8,346,863	8,533,190	97.82%	8,679,288	96.17%
description	14,082,381	16,330,817	86.23%	17,034,777	82.67%

We aligned the extracted data in batches split by section type and year of origin, in order to speed up the process by parallelisation. Each extracted text section was considered to be a document. Table 3 shows the number of sentences on the source (de) and target (en) side after sentence splitting, compared to the number of aligned sentences that were output by Gargantua. The text from the abstract sections exhibits a strong asymmetry, with about 780,000 sentences on the German corresponding to 940,000 sentences on the English side. Given that extremely unbalanced sections were already discarded in the extraction process, this indicates that abstracts are often not literal translations.This asymmetry also results in a lower input to output ratio of 92.36% in relation to the number of possible alignments, which is given by the number of source sentences. The claims are more balanced with 8.5 million sentences on the source and 8.6 million sentences on the target side and exhibit a high input to output ratio of 97.82%. This is presumably owing to the nature of patent claims: mapped out as a numbered list of sentences, they tend to be translated phrase by phrase. Like the abstracts, the descriptions are quite imbalanced and yield the worst input to output ratio with 86.00% on the source side. Furthermore, spot tests show that the aligned data contains some sentence pairs that only partially overlap.

5 Experimental Data

We conduct information-theoretic as well as MT experiments to analyze and characterize the extracted corpus and gain information that indicates possible

[10] http://www.statmt.org/europarl/v6/tools.tgz

[11] A full stop can either indicate the end of a sentence or an abbreviation.

use cases for this data. We split the corpus according to the two previously defined dimensions, namely text sections and top-level IPC classification. In this way, we gain four subcorpora for the structural dimension and eight for the topical, resulting in 24 possible combinations. We furthermore split the data by year of publication date and assign documents published between 1979 and 2007 for training and documents published in 2008 for tuning and testing. To ensure that there is no biasing overlap between training and test data, we compute the percentage of sentences in the extracted data that are actually unique (table 4). Abstract, claims and descriptions are unproblematic with a small amount of duplicates that are partly caused by noise from incorrectly split sentences and partly by sentences that are used in several patents. For instance, the patent documents EP 1050190 (A1), titled Active acoustic devices comprising panel members, and EP 1414266 (A2), titled Active acoustic devices, share four claims word for word. Both patent applications were filed by the same company and name the same inventor. EP 1050190 (A1) was published in 2000, EP 1414266 (A2) four years later. They are obviously related and we can assume that part of the text has been recycled to save writing time, but there is no connection through citation or family id. One could try to prevent such duplicates by exploiting company and inventor metadata, but we argue that these duplicates are a natural characteristic of patent documents and should therefore be kept in the data, if the effect is not biasing evaluation.

For titles, however, the number of duplicates is considerably higher and originates from a different source. Since the title must be short and may only contain technical terms to describe the invention, many patents – from different years and by different companies and inventors – share the same title. For instance, there are 53 patents titled Push button switch in the EP corpus. To prevent a bias, we removed the overlap between test and training set for titles.

Table 4. Percentage of unique sentences relative to the total number of sentences in the experimental data.

title	75.09%
abstract	93.92%
description	96.58%
claims	94.65%

We create the actual training, development and test sets by sampling from the subdomains. From every text section subcorpus we sample 500,000 sentences – distributed across all IPC sections – for training and 2,000 sentences each for development and test set. Table 5 gives the number of types and tokens in the resulting training sets. Note that the title set contains considerably fewer tokens than the other sections – only about 15% of the amount contained in the abstracts – so models trained on this section might suffer from the disadvantage of having seen less data than the others. However, the titles still contain a large

amount of types – more than half as many as the abstracts – so the disadvantage should not weight too heavily. A potentially hurtful property of the German data is the high type-token-ratio, which is caused by German compounding, especially for titles. We therefore investigate the influence of compound splitting on the translation performance on text sections in an additional experiment.

Table 5. Types and tokens on 500k sentences text section training sets.

	de			en		
	#tokens	#types	$\frac{types}{tokens}$	#tokens	#types	$\frac{types}{tokens}$
title	3,267,802	512,773	0.1569	4,038,743	176,293	0.0437
abstract	18,627,983	921,486	0.0495	21,245,542	269,803	0.0127
description	12,836,238	684,190	0.0533	15,961,246	281,053	0.0176
claims	15,646,621	784,978	0.0502	18,355,584	270,013	0.0147

Table 6. Number of sentences per IPC section on claims.

A	1,947,542
B	2,522,995
C	2,263,375
D	299,742
E	353,910
F	1,012,808
G	2,066,132
H	1,754,573

For training the IPC domain models, we chose only sentences from the claims and abstract domain, since this data is possibly the cleanest and generally used in IPC cross-domain experiments. We sampled 300,000 sentences from the training corpus for each IPC section. This is the largest training set possible, since the smallest section, D, (see table 6 for the distribution of IPC sections across the claims) contains overall just barely 300,000 sentences in the combined EP and WO training set[12]. [7] use only the five largest sections of the IPC for their experiments, but since we would like to gain a comprehensive view on the data we include all eight IPC domains in our experiments. The resulting number of types and tokens are shown in table 7. There is less variance than on the text sections; still, we can note that section C contains the largest amount of types, which is likely due to the high number of formulae in this section.

[12] To address the problem of patent duplicates across different corpora, we only include sentences from the WO which came from documents that do not share a family id with a document in EP.

Table 7. Types and tokens in 300k sentences IPC section training sets.

	de			en		
	#tokens	#types	$\frac{types}{tokens}$	#tokens	#types	$\frac{types}{tokens}$
A	9,843,156	520,839	0.0529	11,242,459	233,266	0.0207
B	10,726,633	508,943	0.0474	12,561,139	151,601	0.0121
C	9,514,203	527,609	0.0555	10,942,622	256,932	0.0235
D	13,900,065	440,014	0.0317	16,146,597	160,445	0.0099
E	10,922,892	355,606	0.0326	12,835,170	99,915	0.0078
F	10,941,342	416,113	0.0380	12,941,777	113,498	0.0088
G	10,943,693	578,183	0.0528	12,700,396	180,536	0.0142
H	11,064,367	545,433	0.0493	12,940,731	157,507	0.0122

6 Textual Similarities across IPC Domains

The IPC domains are less well characterized by type-token-ratio, but we expect them to differ with regard to lexical content. To analyize these domains, we therefore compute three information-theoretic similarity measures that perform a pairwise comparison of the vocabulary probability distribution of each task-specific corpus. This distribution is calculated on the basis of the 500 most frequent words in the union of two corpora, normalized by vocabulary size. The first measure is a computation of **Spearman's rank correlation** ([12]) on the frequency-ranked word lists of two corpora. The second measure is a calculation of the **cross-entropy** between the language model probabilities of a model trained on corpus A when applied to corpus B ([13]). As a third metric we use the **$\mathcal{A}$-distance** measure of [14]. If $\mathcal{A}$ is the set of measurable subsets on which the word distributions are defined, then the $\mathcal{A}$-distance is the probability of the subset on which the distributions differ most. A low distance translates to higher similarity.

The three measures for corpus similarity based on the corpus vocabulary are displayed in tables 8, 9 and 10. A low cross-entropy and distance and a high correlation close to 1 or -1 indicate similarity between the vocabulary of two sections. The most similar section or sections – apart from the section itself on the diagonal, highlighted in *italic* font – is indicated in **bold** face. All three measures support a pairwise similarity of A and C, B and F, G and H. Furthermore, a close similarity between E and F is indicated. G and H (electricity and physics, respectively) are very similar to each other but not close to any other section apart from B. This makes sense intuitively, since physics and electricity probably play a more important role in transportations and constructions than, for example, chemistry. Semantically, the two fields are closely related; in fact, the latter can be viewed as a subfield of the former.

The cross-entropy also gives a measure for the homogeneity of each IPC domain on the diagonal; a low perplexity of the language model on a test set from the same domain corresponds to high homogeneity. According to this, C is the most homogeneous domain, followed by A, which is interesting, since these

domains have a comparatively large vocabulary. This may indicate that the cross-entropy is not the best measure to compare different domains.

Table 8. Pairwise $\mathcal{A}$-distance for 300k IPC training sets.

	A	B	C	D	E	F	G	H
A	*0*	0.1303	**0.1317**	0.1311	0.188	0.186	0.164	0.1906
B	**0.1302**	*0*	0.2388	**0.1242**	**0.0974**	**0.0875**	0.1417	0.1514
C	**0.1317**	0.2388	*0*	0.1992	0.311	0.3068	0.2506	0.2825
D	**0.1311**	0.1242	0.1992	*0*	0.1811	0.1808	0.1876	0.201
E	0.188	0.0974	0.311	0.1811	*0*	**0.0921**	0.2058	0.2025
F	0.186	**0.0875**	0.3068	0.1808	**0.0921**	*0*	0.1824	0.1743
G	0.164	0.1417	0.2506	0.1876	0.2056	0.1824	*0*	**0.064**
H	0.1906	0.1514	0.2825	0.201	0.2025	0.1743	**0.064**	*0*

Table 9. Pairwise Spearman's rank correlation for 300k IPC training sets.

	A	B	C	D	E	F	G	H
A	*1*	0.5335	**0.5372**	0.5067	0.333	0.3293	0.3192	0.2093
B	**0.5333**	*1*	0.1539	**0.5496**	**0.6132**	**0.6618**	0.4476	0.366
C	**0.5373**	0.1539	*1*	0.3338	-0.0719	-0.0539	0.0756	-0.0226
D	0.5067	0.5496	0.3337	*1*	0.2585	0.2648	0.2636	0.1645
E	0.3329	0.6131	-0.0719	0.2585	*1*	0.6027	0.1928	0.1933
F	0.3293	**0.6618**	-0.0539	0.2648	**0.6027**	*1*	0.2645	0.2684
G	0.319	0.4477	0.0756	0.2636	0.1936	0.2646	*1*	**0.751**
H	0.2091	0.3661	-0.0226	0.1645	0.1933	0.2683	**0.7509**	*1*

7 Cross-Domain Translation

We conducted a first investigation of the performance of our corpus in MT. We view these experiments as an expansion of the similarity analysis and therefore only look at cross-domain evaluation. We deliberately do not compare the domain-specific translation models to a model pooled from all data, since previous work has already shown that this outperforms smaller individual models.

We used the phrase-based, open-source SMT toolkit Moses[13] [15] with the standard feature set. We computed 5-gram language models on the target side of the training set with IRSTLM[14] [16] and queried the model with KenLM [17].

[13] `http://statmt.org/moses/`

[14] `http://sourceforge.net/projects/irstlm/`

Table 10. Cross-entropy: pairwise 300k language model perplexity on 2k IPC test set.

	test							
train	A	B	C	D	E	F	G	H
A	*210.8*	320.5	**275.7**	413.4	438.4	417.2	295.6	374.9
B	260.3	*230.0*	332.0	346.5	329.6	**293.5**	**273.3**	**294.6**
C	**220.8**	308.8	*181.0*	**362.6**	490.3	456.6	**281.5**	378.2
D	252.3	248.6	281.1	*239.8*	356.1	332.8	303.3	332.7
E	267.0	254.3	375.9	374.3	*258.0*	**290.3**	299.7	312.9
F	267.1	**239.3**	355.3	**361.0**	**308.6**	*233.8*	**277.9**	**288.1**
G	360.6	367.1	472.9	565.3	539.3	467.5	*255.7*	303.5
H	402.7	386.5	529.5	606.8	546.4	469.5	**280.7**	*257.7*

The cross-domain evaluation[15] on the IPC classes (table 11) shows that every subdomain is best translated with a model trained on the respective section: the BLEU scores on the diagonal are the highest in every column. For assessing similarities, we are therefore interested in the runner-up on each section (indicated in **bold** font). Note that best section scores vary considerably, ranging from 0.5719 on C to 0.4714 on H, indicating classes that are easier to translate. C, the Chemistry section, presumably benefits from the fact that the data contains chemical formulae, which are language-independent and do not have to be translated. [7] show similar variations on the five largest IPC classes with scores ranging from 0.609 on C to 0.5518 on G for the PLuTO data. The higher overall scores and smaller variance are due to the larger amounts of training data used in these experiments; we have opted for a reduced training set in favour of including the smaller sections D, E, and F in our experiments, which were omitted from the PLuTO experiments. Again, for determining the relationship between the domains, we examine the best runner-up on each section, considering the BLEU score, although asymmetrical, as a kind of measure of similarity between domains. We can establish symmetric relationships between sections A and C, B and F as well as G and H, which means that the models are mutual runner-up on the other's test section. This shows that the same relationships can be inferred from BLEU scores as from the information theoretic measures evaluated before.

To ensure that these effects are not solely caused by the overlap between IPC classes, table 12 shows the relative pairwise document overlap between IPC sections, i.e. the percentage of documents in A that are also classified as B, C, D etc. in column A. We observe a large overlap of roughly 30% in both directions between A and C, accounting for the mutual runner-up result in the cross-section evaluation; the same holds for G and H. However, the amount of overlap is not the sole factor for the mutual translation performance on sections. C and B share over 20% of their documents, but C performs worst on B and vice versa. The relationship between sections is also not necessarily symmetrical: the smaller sections D, E, and F each share about 30% of their documents with B,

[15] We computed BLEU_4 [18] on lowercased data.

Table 11. BLEU scores for 300k individual IPC section models.

	test							
train	A	B	C	D	E	F	G	H
A	*0.5349*	0.4475	**0.5472**	0.4746	0.4438	0.4523	0.4318	0.4109
B	0.4846	*0.4736*	0.5161	**0.4847**	**0.4578**	**0.4734**	0.4396	0.4248
C	**0.5047**	0.4257	*0.5719*	0.462	0.4134	0.4249	0.409	0.3845
D	0.47	0.4387	0.5106	*0.5167*	0.4344	0.4435	0.407	0.3917
E	0.4486	0.4458	0.4681	0.4531	*0.4771*	0.4591	0.4073	0.4028
F	0.4595	**0.4588**	0.4761	0.4655	0.4517	*0.4909*	0.422	0.4188
G	0.4935	0.4489	0.5239	0.4629	0.4414	0.4565	*0.4748*	**0.4532**
H	0.4628	0.4484	0.4914	0.4621	0.4421	0.4616	**0.4588**	*0.4714*

which is mirrored in the translation score of B on all three sections where it is always runner-up. In the other direction, the influcence of these sections on B is only small, but there seems to be a strong similarity between B and F which is significantly stronger than the relationship of B to the other small sections. We conclude that the document overlap between two sections is an indicator but not the determining parameter for similarity and mutual translation performance.

Table 12. Percentage document IPC overlap: $\frac{|X \cap Y|}{|Y|}$.

	Y							
X	A	B	C	D	E	F	G	H
A	*100.0%*	9.7%	**31.5%**	15.4%	6.0%	5.9%	10.0%	2.5%
B	13.0%	*100.0%*	24.5%	**32.0%**	**23.6%**	**30.3%**	17.3%	11.7%
C	**36.2%**	**21.0%**	*100.0%*	28.4%	7.8%	8.2%	14.4%	8.9%
D	2.4%	3.8%	3.9%	*100.0%*	1.7%	1.5%	0.7%	0.5%
E	1.2%	3.5%	1.3%	2.1%	*100.0%*	5.0%	1.4%	0.9%
F	3.3%	12.6%	4.0%	5.2%	14.2%	*100.0%*	4.6%	5.0%
G	10.7%	13.8%	13.4%	5.0%	7.9%	8.9%	*100.0%*	**30.3%**
H	2.3%	8.0%	7.1%	2.9%	4.1%	8.3%	**26.0%**	*100.0%*

Evaluation results for separately trained individual models across text domains are shown in table 13 and exhibit patterns similar to the evaluation of the IPC models. Again, each section is best translated with a model trained on data from the same section. The results on abstracts suggest that this section most strongly resembles the claims; the model trained on claims achieves a respectable score. On claims, abstract and description models yield an almost equal score, but the score drops substantially from the best result, supporting the notion that claims possess a very distinct structure and wording that is only captured by a model that is able to learn these characteristics from the data. With this data available, however, claims seem to be easiest to translate, yielding

the highest overall BLEU score of 0.4879. On the other hand, all models score considerably lower on title data, which is no surprise considering the fact that titles consist only of noun phrases and translation quality depends highly on vocabulary coverage. Given the high type-token-ratio of this section, we discuss a method to expand coverage in the next section. The parallel data obtained from the descriptions presumably lacks in quality compared to the other sections due to its origin. Overall, the scores on descriptions are lower than on abstracts and claims but still higher than on titles. The abstract model again scores best on this section. Altogether, the abstract model seems to be the most robust and varied model, yielding the runner-up score on all other sections. The title model, in contrast, performs worst across all other sections. We attribute these results to the limited variety of grammatical structure observed in the training data – titles only consist of noun phrases – as well as the smaller amount of training data with regard to the absolute number of tokens.

Table 13. BLEU scores for 500k text domain models.

	test			
train	title	abstract	description	claims
title	*0.3196*	0.2839	0.1743	0.3512
abstract	**0.2681**	*0.3737*	**0.2812**	**0.4076**
description	0.2342	0.32189	*0.3347*	**0.403**
claims	0.2623	**0.3416**	0.2420	*0.4879*

8 Compound Splitting on Textual Domains

We hypothesize that the high type-token-ratio on the German source side stems from a large number of compound words, due to the fact that inventions have to be described accurately and no proper names may be used. We therefore investigated the effect of German compound splitting as a preprocessing step on the text section dimensions. We trained and applied a compound splitting model on the German part of the training sets for abstract, claims and titles and the respective test sets. Predictably, the splitting raises the token count and average type frequency while lowering the type count.

We apply a simple empirical compound splitting method by [19], which is distributed as a script in the Moses toolkit. It considers all possible splittings of a compound word into known words, taking possible fillers, such as the `s` in Rotationsverdämpfer, and dropped letters into account. The vocabulary of known words is derived from a monolingual training corpus. The decision, if and how a word will be split, is based on word frequency estimates made on the same corpus. Given the word count in the corpus, the model picks the split S with the highest geometric mean of word frequencies of its parts p_i: $\text{argmax}_S(\prod_{p_i \in S} \text{count}(p_i))^{\frac{1}{n}}$.

This approach has the effect that if a compound word appears more frequently than its parts, it is left intact, ensuring that common compounds are not unnecessarily broken up. A phrase-based translation model will learn a correct translation even when an incorrect split is done consistently, if the compounds in the training data are split as well.

The effects of compound splitting for German-English translation are displayed in table 14. Compound splitting improves the score on all four sections but most substantially on the titles with an improvement of 0.0439 BLEU. This shows that the translation of titles is strongly influcenced by a large percentage of compound words and suffers mostly from a sparse data problem. Table 15 shows the influence of compound splitting on two translation samples. The splitting does not always result in a perfect translation, e.g. producing air jet instead of air nozzle, but it considerably reduces the out-of-vocabulary (OOV) rate.

Table 14. Compound splitting on training and test set.

	BLEU score		OOV rate	
	raw	split	raw	split
title	0.3196	**0.3635**	10.00%	5.44%
abstract	0.3737	**0.3827**	3.35%	1.20%
description	0.3347	**0.3385**	3.90%	1.60%
claims	0.4879	**0.5022**	4.58%	2.42%

Table 15. Sample compound splitting on titles.

source	Luftdüsenspinnmaschinen	Druckluftversorgungssystems
split source	luft düsen spinnmaschinen	druck luft versorgung systems
baseline system	OOV	OOV
compound split system	air jet spinning machine	compressed air supply system
reference	air nozzle spinning machine	compressed air supply system

9 Discussion

Statistical machine translation is highly dependant on the availability of sentence-parallel data for diverse domains. We documented a twofold approach to extract large amounts of parallel data from the MAREC patent corpus, namely a straight-forward method, where we align translated document sections, i.e. title, abstract and claims, and an indirect approach, where we find approximate translations of a whole document via the patent family id connection. The full

statistics for the resulting German-English parallel corpus can be found in table 16, containing the number of unique parallel sentences and number of tokens on English and German side. The large amount of clean, parallel data consitutes a valuable resource for patent translation. Based on this, we plan to explore advanced topics such as cross-lingual patent retrieval.

Table 16. Full sentence-parallel corpus: number of unique sentences, number of tokens.

	sentences	tokens en	tokens de
title	2,204,384	18,159,477	14,798,306
abstract	715,735	30,830,602	26,606,213
description	11,912,840	446,229,748	322,282,966
claims	8,181,791	496,421,795	431,568,084
total	23,014,750	991,641,622	795,255,569

We further analysed the corpus by exploring two different subdomain dimensions in terms of their relatedness, both with corpus similarits measures and machine translation evaluation. We find that both the IPC and the document structure domains are well-delimitated and showed that pairwise domain similarity and translation performence correlate. Furthermore, we identify and adress one particular structural problem for German-English patent translation, namely the high type-token-ratio, which stems from the large amount of new terms produced by German compounding.

We plan a similar extraction of parallel data for French-English and French-German patent translation in the future. These resources will enable us to further identify and investigate topics in automated patent translation.

Acknowledgements

The authors would like to thank the Information Retrieval Facility (IRF) for providing the MAREC patent corpus.

This work was supported in part by DFG grant "Cross-language Learning-to-Rank for Patent Retrieval".

References

1. Wäschle, K., Riezler, S.: Structural and topical dimensions in multi-task patent translation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France (2012)
2. Utiyama, M., Isahara, H.: A japanese-english patent parallel corpus. In: Proceedings of MT Summit XI, Copenhagen, Denmark (2007)
3. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Computational Linguistics **19**(1) (1993) 75–102

4. Lu, B., Tsou, B.K., Zhu, J., Jiang, T., Kwong, O.Y.: The construction of a chinese-english patent parallel corpus. In: Proceedings of the MT Summit XII, Ottawa, Canada (2009)
5. Tinsley, J., Way, A., Sheridan, P.: PLuTO: MT for online patent translation. In: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, CO (2010)
6. Jochim, C., Lioma, C., Schütze, H., Koch, S., Ertl, T.: Preliminary study into query translation for patent retrieval. In: Proceedings of the 3rd International Workshop on Patent Information Retrieval (PaIR 2010), Toronto, Canada (2010)
7. Ceauşu, A., Tinsley, J., Zhang, J., Way, A.: Experiments on domain adaptation for patent machine translation in the PLuTO project. In: Proceedings of the 15th Conference of the European Assocation for Machine Translation (EAMT 2011), Leuven, Belgium (2011)
8. Braune, F., Fraser, A.: Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10), Beijing, China (2010)
9. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
10. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X, Phuket, Thailand (2005)
11. Moore, R.: Fast and accurate sentence alignment of bilingual corpora. In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA'02), Tiburon, CA (2002)
12. Siegel, S., Castellan, J.: Nonparametric Statistics for the Behavioral Sciences. Second Edition. MacGraw-Hill, Boston, MA (1988)
13. Kilgarriff, A., Rose, T.: Measures for corpus similarity and homogeneity. In: Proceedings of the 3rd conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain (1998)
14. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS'06), Vancouver, Canada (2006)
15. Koehn, P., Hoang, H., Birch, A., Callison-Birch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Czech Republic (2007)
16. Federico, M., Bertoldi, N., Cettolo, M.: IRSTLM: an open source toolkit for handling large scale language models. In: Proceedings of Interspeech, Brisbane, Australia (2008)
17. Heafield, K.: KenLN: faster and smaller language model queries. In: Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation (WMT'11), Edinburgh, UK (2011)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Technical Report IBM Research Division Technical Report, RC22176 (W0190-022), Yorktown Heights, N.Y. (2001)
19. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics (EACL'03), Budapest, Hungary (2003)