The Impact of Spelling Errors on Patent Search

Benno Stein and **Dennis Hoppe** and **Tim Gollub**

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract

The search in patent databases is a risky business compared to the search in other domains. A single document that is relevant but overlooked during a patent search can turn into an expensive proposition. While recent research engages in specialized models and algorithms to improve the effectiveness of patent retrieval, we bring another aspect into focus: the detection and exploitation of patent inconsistencies. In particular, we analyze spelling errors in the assignee field of patents granted by the United States Patent & Trademark Office. We introduce technology in order to improve retrieval effectiveness despite the presence of typographical ambiguities. In this regard, we (1) quantify spelling errors in terms of edit distance and phonological dissimilarity and (2) render error detection as a learning problem that combines word dissimilarities with patent meta-features. For the task of finding all patents of a company, our approach improves recall from 96.7% (when using a state-of-the-art patent search engine) to 99.5%, while precision is compromised by only 3.7%.

1 Introduction

Patent search forms the heart of most retrieval tasks in the intellectual property domain—cf. Table 1, which provides an overview of various user groups along with their typical (•) and related (○) tasks. The due diligence task, for example, is concerned with legal issues that arise while investigating another company. Part of an investigation is a patent portfolio comparison between one or more competitors (Lupu et al., 2011). Within all tasks recall is preferred over precision, a fact which distinguishes patent search from general web search. This retrieval constraint has produced a variety of sophisticated approaches tailored to the patent domain: citation analysis (Magdy and Jones, 2010), the learning of section-specific retrieval models (Lopez and Romary, 2010), and automated query generation (Xue and Croft, 2009). Each approach improves retrieval performance, but what keeps them from attaining maximum effectiveness in terms of recall are the inconsistencies found in patents: incomplete citation sets, incorrectly assigned classification codes, and, not least, spelling errors.

Our paper deals with spelling errors in an obligatory and important field of each patent, namely, the patent assignee name. Bibliographic fields are widely used among professional patent searchers in order to constrain keyword-based search sessions (Joho et al., 2010). The assignee name is particularly helpful for patentability searches and portfolio analyses since it determines the company holding the patent. Patent experts address these search tasks by formulating queries containing the company name in question, in the hope of finding all patents owned by that company. A formal and more precise description of this relevant search task is as follows: Given a query q which specifies a company, and a set D of patents, determine the set $D_q \subset D$ comprised of all patents held by the respective company.

For this purpose, all assignee names in the patents in D should be analyzed. Let A denote the set of all assignee names in D, and let $a \sim q$ denote the fact that an assignee name $a \in A$ refers to company q. Then in the portfolio search task, all patents filed under a are relevant. The retrieval of D_q can thus be rendered as a query expansion

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 570–579,
Avignon, France, April 23 - 27 2012.

Table 1: User groups and patent-search-related retrieval tasks in the patent domain (Hunt et al., 2007).

		User group					
		Analyst	Attorney	Manager	Inventor	Investor	Researcher
Patent search task	Patentability	●	○		●		○
	State of the art					○	●
	Infringement		●				
	Opposition		●			●	
	Due diligence		●	●			
	Portfolio	●	○	●		●	

task, where q is expanded by the disjunction of assignee names A_q with $A_q = \{a \in A \mid a \sim q\}$.

While the trivial expansion of q by the entire set A ensures maximum recall but entails an unacceptable precision, the expansion of q by the empty set yields a reasonable baseline. The latter approach is implemented in patent search engines such as PatBase[1] or FreePatentsOnline,[2] which return all patents where the company name q occurs as a substring of the assignee name a. This baseline is simple but reasonable; due to trademark law, a company name q must be a unique identifier (i.e. a key), and an assignee name a that contains q can be considered as relevant. It should be noted in this regard that $|q| < |a|$ holds for most elements in A_q, since the assignee names often contain company suffixes such as "Ltd" or "Inc".

Our hypothesis is that due to misspelled assignee names a substantial fraction of relevant patents cannot be found by the baseline approach. In this regard, the types of spelling errors in assignee names given in Table 2 should be considered.

Table 2: Types of spelling errors with increasing problem complexity according to Stein and Curatolo (2006). The first row refers to lexical errors, whereas the last two rows refer to phonological errors. For each type, an example is given, where a misspelled company name is followed by the correctly spelled variant.

Spelling error type	**Example**
Permutations or dropped letters	Whirpool Corporation → Whirlpool Corporation
Misremembering spelling details	Whetherford International → Weatherford International
Spelling out the pronunciation	Emulecks Corporation → Emulex Corporation

In order to raise the recall for portfolio search without significantly impairing precision, an approach more sophisticated than the standard retrieval approach, which is the expansion of q by the empty set, is needed. Such an approach must strive for an expansion of q by a subset of A_q, whereby this subset should be as large as possible.

1.1 Contributions

The paper provides a new solution to the problem outlined. This solution employs machine learning on orthographic features, as well as on patent meta features, to reliably detect spelling errors. It consists of two steps: (1) the computation of A_q^+, the set of assignee names that are in a certain edit distance neighborhood to q; and (2) the filtering of A_q^+, yielding the set A_q^*, which contains those assignee names from A_q^+ that are classified as misspellings of q. The power of our approach can be seen from Table 3, which also shows a key result of our research; a retrieval system that exploits our classifier will miss only 0.5% of the relevant patents, while retrieval precision is compromised by only 3.7%.

Another contribution relates to a new, manually-labeled corpus comprising spelling errors in the assignee field of patents (cf. Section 3). In this regard, we consider the over 2 million patents granted by the USPTO between 2001 and 2010. Last, we analyze indications of deliberately inserted spelling errors (cf. Section 4).

Table 3: Mean average Precision, Recall, and F-Measure ($\beta = 2$) for different expansion sets for q in a portfolio search task, which is conducted on our test corpus (cf. Section 3).

Expansion set for q		**Precision**	**Recall**	$\mathbf{F_2}$
$\emptyset$	(baseline)	**0.993**	0.967	0.968
A_q^*	(machine learning)	0.956	**0.995**	0.980
A	(trivial)	0.001	1.0	0.005
A_q^+	(edit distance)	0.274	1.0	0.672

[1] www.patbase.com

[2] www.freepatentsonline.com

1.2 Causes for Inconsistencies in Patents

We identify the following six factors for inconsistencies in the bibliographic fields of patents, in particular for assignee names: (1) Misspellings are introduced due to the lack of knowledge, the lack of attention, and due to spelling disabilities. Intellevate Inc. (2006) reports that 98% of a sample of patents taken from the USPTO database contain errors, most which are spelling errors. (2) Spelling errors are only removed by the USPTO upon request (U.S. Patent & Trademark Office, 2010). (3) Spelling variations of inventor names are permitted by the USPTO. The Manual of Patent Examining Procedure (MPEP) states in paragraph 605.04(b) that "if the applicant's full name is 'John Paul Doe,' either 'John P. Doe' or 'J. Paul Doe' is acceptable." Thus, it is valid to introduce many different variations: with and without initials, with and without a middle name, or with and without suffixes. This convention applies to assignee names, too. (4) Companies often have branches in different countries, where each branch has its own company suffix, e.g., "Limited" (United States), "GmbH" (Germany), or "Kabushiki Kaisha" (Japan). Moreover, the usage of punctuation varies along company suffix abbreviations: "L.L.C." in contrast to "LLC", for example. (5) Indexing errors emerge from OCR processing patent applications, because similar looking letters such as "e" versus "c" or "l" versus "I" are likely to be misinterpreted. (6) With the advent of electronic patent application filing, the number of patent reexamination steps was reduced. As a consequence, the chance of undetected spelling errors increases (Adams, 2010).

All of the mentioned factors add to a highly inconsistent USPTO corpus.

2 Related Work

Information within a corpus can only be retrieved effectively if the data is both accurate and unique (Müller and Freytag, 2003). In order to yield data that is accurate and unique, approaches to data cleansing can be utilized to identify and remove inconsistencies. Müller and Freytag (2003) classify inconsistencies, where duplicates of entities in a corpus are part of a semantic anomaly. These duplicates exist in a database if two or more different tuples refer to the same entity. With respect to the bibliographic fields of patents, the assignee names "Howlett-Packard" and "Hewett-Packard" are distinct but refer to the same company. These kinds of near-duplicates impede the identification of duplicates (Naumann and Herschel, 2010).

Near-duplicate Detection The problem of identifying near-duplicates is also known as record linkage, or name matching; it is subject of active research (Elmagarmid et al., 2007). With respect to text documents, slightly modified passages in these documents can be identified using fingerprints (Potthast and Stein, 2008). On the other hand, for data fields which contain natural language such as the assignee name field, string similarity metrics (Cohen et al., 2003) as well as spelling correction technology are exploited (Damerau, 1964; Monge and Elkan, 1997). String similarity metrics compute a numeric value to capture the similarity of two strings. Spelling correction algorithms, by contrast, capture the likelihood for a given word being a misspelling of another word. In our analysis, the similarity metric *SoftTfIdf* is applied, which performs best in name matching tasks (Cohen et al., 2003), as well as the complete range of spelling correction algorithms shown in Figure 1: Soundex, which relies on similarity hashing (Knuth, 1997), the Levenshtein distance, which gives the minimum number of edits needed to transform a word into another word (Levenshtein, 1966), and SmartSpell, a phonetic production approach that computes the likelihood of a misspelling (Stein and Curatolo, 2006). In order to combine the strength of multiple metrics within a near-duplicate detection task, several authors resort to machine learning (Bilenko and Mooney, 2002; Cohen et al., 2003). Christen (2006) concludes that it is important to exploit all kinds of knowledge about the type of data in question, and that inconsistencies are domain-specific. Hence, an effective near-duplicate detection approach should employ domain-specific heuristics and algorithms (Müller and Freytag, 2003). Following this argumentation, we augment various word similarity assessments with patent-specific meta-features.

Patent Search Commercial patent search engines, such as PatBase and FreePatentsOnline, handle near-duplicates in assignee names as follows. For queries which contain a company name followed by a wildcard operator, PatBase suggests

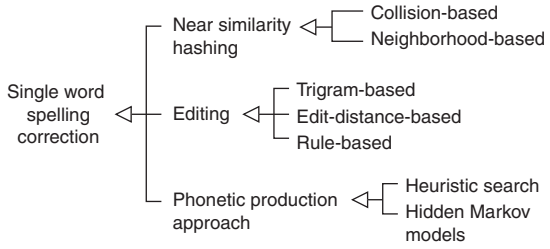

Figure 1: Classification of spelling correction methods according to Stein and Curatolo (2006).

a set of additional companies (near-duplicates), which can be considered alongside the company name in question. These suggestions are solely retrieved based on a trailing wildcard query. Each additional company name can then be marked individually by a user to expand the original query. In case the entire set of suggestions is considered, this strategy conforms to the expansion of a query by the empty set, which equals a reasonable baseline approach. This query expansion strategy, however, has the following drawbacks: (1) The strategy captures only inconsistencies that succeed the given company name in the original query. Thus, near-duplicates which contain spelling errors in the company name itself are not found. Even if PatBase would support left trailing wildcards, then only the full combination of wildcard expressions would cover all possible cases of misspellings. (2) Given an acronym of a company such as IBM, it is infeasible to expand the abbreviation to "International Business Machines" without considering domain knowledge.

Query Expansion Methods for Patent Search To date, various studies have investigated query expansion techniques in the patent domain that focus on prior-art search and invalidity search (Magdy and Jones, 2011). Since we are dealing with queries that comprise only a company name, existing methods cannot be applied. Instead, the near-duplicate task in question is more related to a text reuse detection task discussed by Hagen and Stein (2011); given a document, passages which also appear identical or slightly modified in other documents, have to be retrieved by using standard keyword-based search engines. Their approach is guided by the user-over-ranking hypothesis introduced by Stein and Hagen (2011). It states that "the best retrieval performance can be achieved with queries returning about as many results as can be considered at user site." If we make use of their terminology, then we can distinguish the query expansion sets (cf. Table 3) into two categories: (1) The trivial as well as the edit distance expansion sets are *underspecific*, i.e., users cannot cope with the large amount of irrelevant patents returned; the precision is close to zero. (2) The baseline approach, by contrast, is *overspecific*; it returns too few documents, i.e., the achieved recall is not optimal. As a consequence, these query expansion sets are not suitable for portfolio search. Our approach, on the other hand, excels in both precision and recall.

Query Spelling Correction Queries which are submitted to standard web search engines differ from queries which are posed to patent search engines with respect to both length and language diversity. Hence, research in the field of web search is concerned with suggesting reasonable alternatives to misspelled queries rather than correcting single words (Li et al., 2011). Since standard spelling correction dictionaries (e.g. ASpell) are not able to capture the rich language used in web queries, large-scale knowledge sources such as Wikipedia (Li et al., 2011), query logs (Chen et al., 2007), and large n-gram corpora (Brants et al., 2007) are employed. It should be noted that the set of correctly written assignee names is unknown for the USPTO patent corpus.

Moreover, spelling errors are modeled on the basis of language models (Li et al., 2011). Okuno (2011) proposes a generative model to encounter spelling errors, where the original query is expanded based on alternatives produced by a small edit distance to the original query. This strategy correlates to the trivial query expansion set (cf. Section 1). Unlike using a small edit distance, we allow a reasonable high edit distance to maximize the recall.

Trademark Search The trademark search is about identifying registered trademarks which are similar to a new trademark application. Similarities between trademarks are assessed based on figurative and verbal criteria. In the former case, the focus is on image-based retrieval techniques. Trademarks are considered verbally similar for a variety of reasons, such as pronunciation, spelling, and conceptual closeness, e.g., swapping letters or using numbers for words. The verbal similarity of trademarks, on the other hand, can be determined by using techniques comparable to near-duplicate detection: phonological parsing,

fuzzy search, and edit distance computation (Fall and Giraud-Carrier, 2005).

3 Detection of Spelling Errors

This section presents our machine learning approach to expand a company query q; the classifier c delivers the set $A_q^* = \{a \in A \mid c(q,a) = 1\}$, an approximation of the ideal set of relevant assignee names A_q. As a classification technology a support vector machine with linear kernel is used, which receives each pair (q,a) as a six-dimensional feature vector. For training and test purposes we identified misspellings for 100 different company names. A detailed description of the constructed test corpus and a report on the classifiers performance is given in the remainder of this section.

3.1 Feature Set

The feature set comprises six features, three of them being orthographic similarity metrics, which are computed for every pair (q,a). Each metric compares a given company name q with the first $|q|$ words of the assignee name a:

1. *SoftTfIdf.* The SoftTfIdf metric is considered, since the metric is suitable for the comparison of names (Cohen et al., 2003). The metric incorporates the Jaro-Winkler metric (Winkler, 1999) with a distance threshold of 0.9. The frequency values for the similarity computation are trained on A.
2. *Soundex.* The Soundex spelling correction algorithm captures phonetic errors. Since the algorithm computes hash values for both q and a, the feature is 1 if these hash values are equal, 0 otherwise.
3. *Levenshtein distance.* The Levenshtein distance for (q,a) is normalized by the character length of q.

To obtain further evidence for a misspelling in an assignee name, meta information about the patents in D, to which the assignee name refers to, is exploited. In this regard, the following three features are derived:

1. *Assignee Name Frequency.* The number of patents filed under an assignee name a: $F_{Freq}(a) = Freq(a, D)$. We assume that the probability of a misspelling to occur multiple times is low, and thus an assignee name with a misspelled company name has a low frequency.
2. *IPC Overlap.* The IPC codes of a patent specify the technological areas it applies to. We assume that patents filed under the same company name are likely to share the same set of IPC codes, regardless whether the company name is misspelled or not. Hence, if we determine the IPC codes of patents which contain q in the assignee name, $IPC(q)$, and the IPC codes of patents filed under assignee name a, $IPC(a)$, then the intersection size of the two sets serves as an indicator for a misspelled company name in a:
$$F_{IPC}(q,a) = \frac{IPC(q) \cap IPC(a)}{IPC(q) \cup IPC(a)}$$
3. *Company Suffix Match.* The suffix match relies on the company suffixes $Suffixes(q)$ that occur in the assignee names of A containing q. Similar to the IPC overlap feature, we argue that if the company suffix of a exists in the set $Suffixes(q)$, a misspelling in a is likely: $F_{Suffixes}(q,a) = 1$ iff $Suffixes(a) \in Suffixes(q)$.

3.2 Webis Patent Retrieval Assignee Corpus

A key contribution of our work is a new corpus called Webis Patent Retrieval Assignee Corpus 2012 (Webis-PRA-12). We compiled the corpus in order to assess the impact of misspelled companies on patent retrieval and the effectiveness of our classifier to detect them.[3] The corpus is built on the basis of 2 132 825 patents D granted by the USPTO between 2001 and 2010; the patent corpus is provided publicly by the USPTO in XML format. Each patent contains bibliographic fields as well as textual information such as the abstract and the claims section. Since we are interested in the assignee name a associated with each patent $d \in D$, we parse each patent and extract the assignee name. This yields the set A of 202 846 different assignee names. Each assignee name refers to a set of patents, which size varies from 1 to 37 202 (the number of patents filed under "International Business Machines Corporation"). It should be noted that for a portfolio

[3] The Webis-PRA-12 corpus is freely available via `www.webis.de/research/corpora`

Table 4: Statistics of spelling errors for the 100 companies in the Webis-PRA-12 corpus. Considered are the number of words and the number of letters in the company names, as well as the number of different company suffixes that are used together with a company name (denoted as variants of q)

	Total	**Num. of words in q**			**Num. of letters in q**			**Num. of variants of q**		
		1	2	3-4	2-10	11-15	16-35	1-5	6-15	16-96
Number of companies in Q	100	36	53	11	30	35	35	45	32	23
Avg. num. of misspellings in A	3.79	2.13	3.75	9.36	1.16	2.94	6.88	0.91	3.81	9.39

search task the number of patents which refer to an assignee name matters for the computation of precision and recall. If we, however, isolate the task of detecting misspelled company names, then it is also reasonable to weight each assignee name equally and independently from the number of patents it refers to. Both scenarios are addressed in the experiments.

Given A, the corpus construction task is to map each assignee name $a \in A$ to the company name q it refers to. This gives for each company name q the set of relevant assignee names A_q. For our corpus, we do not construct A_q for all company names but take a selection of 100 company names from the 2011 Fortune 500 ranking as our set of company names Q. Since the Fortune 500 ranking contains only large companies, the test corpus may appear to be biased towards these companies. However, rather than the company size the structural properties of a company name are determinative; our sample includes short, medium, and long company names, as well as company names with few, medium, and many different company suffixes. Table 4 shows the distribution of company names in Q along these criteria in the first row.

For each company name $q \in Q$, we apply a semi-automated procedure to derive the set of relevant assignee names A_q. In a first step, all assignee names in A which do not refer to the company name q are filtered automatically. From a preliminary evaluation we concluded that the Levenshtein distance $d(q, a)$ with a relative threshold of $|q|/2$ is a reasonable choice for this filtering step. The resulting sets $A_q^+ = \{a \in A \mid d(q, a) \leq |q|/2)$ contain, in total over Q, 14 189 assignee names. These assignee names are annotated by human assessors within a second step to derive the final set A_q for each $q \in Q$. Altogether we identify 1 538 assignee names that refer to the 100 companies in Q. With respect to our classification task, the assignee names in each A_q are positive examples; the remaining assignee names $A_q^+ \setminus A_q$ form the set of negative examples (12 651 in total).

During the manual assessment, names of assignees which include the correct company name q were distinguished from misspelled ones. The latter holds true for 379 of the 1 538 assignee names. These names are not retrievable by the baseline system, and thus form the main target for our classifier. The second row of Table 4 reports on the distribution of the 379 misspelled assignee names. As expectable, the longer the company name, the more spelling errors occur. Companies which file patents under many different assignee names are likelier to have patents with misspellings in the company name.

3.3 Classifier Performance

For the evaluation with the Webis-PRA-12 corpus, we train a support vector machine,[4] which considers the six outlined features, and compare it to the other expansion techniques. For the training phase, we use $2/3$ of the positive examples to form a balanced training set of 1 025 positive and 1 025 negative examples. After 10-fold cross validation, the achieved classification accuracy is 95.97%.

For a comparison of the expansion techniques on the test set, which contains the examples not considered in the training phase, two tasks are distinguished: finding near duplicates in assignee names (cf. Table 5, Columns 3–5), and finding all patents of a company (cf. Table 5, Columns 6–8). The latter refers to the actual task of portfolio search. It can be observed that the performance improvements on both tasks are pretty similar. The baseline expansion $\emptyset$ yields a recall of 0.83 in the first task. The difference of 0.17 to a perfect recall can be addressed by considering query expansion techniques. If the trivial expansion A is applied to the task the maximum recall can be achieved, which, however,

[4] We use the implementation of the WEKA toolkit with default parameters.

Table 5: The search results (macro-averaged) for two retrieval tasks and various expansion techniques. Besides Precision and Recall, the F-Measure with $\beta = 2$ is stated.

Misspelling detection	**Task: assignee names**			**Task: patents**		
	P	R	F_2	P	R	F_2
Baseline ($\emptyset$)	**.975**	.829	.838	**.993**	.967	.968
Trivial (A)	.000	**1.0**	.001	.001	**1.0**	.005
Edit distance (A_q^+)	.274	**1.0**	.499	.412	**1.0**	.672
SVM (Levenshtein)	.752	.981	.853	.851	.991	.911
SVM (SoftTfIdf)	.702	.980	.796	.826	.993	.886
SVM (Soundex)	.433	.931	.624	.629	.984	.759
SVM (orthographic features)	.856	.975	.922	.942	.990	.967
SVM (A_q^*, all features)	.884	.975	**.938**	.956	.995	**.980**

is bought with precision close to zero. Using the edit distance expansion A_q^+ yields a precision of 0.274 while keeping the recall at maximum. Finally, the machine learning expansion A_q^* leads to a dramatic improvement (cf. Table 5, bottom lines), whereas the exploitation of patent meta-features significantly outperforms the exclusive use of orthography-related features; the increase in recall which is achieved by A_q^* is statistically significant (matched pair t-test) for both tasks (assignee names task: $t = -7.6856$, $df = 99$, $p = 0.00$; patents task: $t = -2.1113$, $df = 99$, $p = 0.037$). Note that when being applied as a single feature none of the spelling metrics (Levenshtein, SoftTfIdf, Soundex) is able to achieve a recall close to 1 without significantly impairing the precision.

4 Distribution of Spelling Errors

Encouraged by the promising retrieval results achieved on the Webis-PRA-12 corpus, we extend the analysis of spelling errors in patents to the entire USPTO corpus of granted patents between 2001 and 2010. The analysis focuses on the following two research questions:

1. *Are spelling errors an increasing issue in patents?* According to Adams (2010), the amount of spelling errors should have been increased in the last years due to the electronic patent filing process (cf. Section 1.2). We address this hypothesis by analyzing the distribution of spelling errors in company names that occur in patents granted between 2001 and 2010.

2. *Are misspellings introduced deliberately in patents?* We address this question by analyzing the patents with respect to the eight technological areas based on the International Patent Classification scheme IPC: A (Human necessities), B (Performing operations; transporting), C (Chemistry; metallurgy), D (Textiles; paper), E (Fixed constructions), F (Mechanical engineering; lighting; heating; weapons; blasting), G (Physics), and H (Electricity). If spelling errors are introduced accidentally, then we expect them to be uniformly distributed across all areas. A biased distribution, on the other hand, indicates that errors might be inserted deliberately.

In the following, we compile a second corpus on the basis of the entire set A of assignee names. In order to yield a uniform distribution of the companies across years, technological areas and countries, a set of 120 assignee names is extracted for each dimension. After the removal of duplicates, we revised these assignee names manually in order to check (and correct) their spelling. Finally, trailing business suffixes are removed, which results in a set of 3 110 company names. For each company name q, we generate the set A_q^* as described in Section 3.

The results of our analysis are shown in Table 6. Table 6(a) refers to the first research question and shows that the amount of misspellings in companies decreased over the years from 6.67% in 2001 to 4.74% in 2010 (cf. Row 3). These results let us reject the hypothesis of Adams (2010). Nevertheless, the analysis provides evidence that spelling errors are still an issue. For example, the company identified with most spelling errors are "Koninklijke Philips Electronics" with 45 misspellings in 2008, and "Centre National de la Recherche Scientifique" with 28 misspellings in 2009. The results are consistent with our findings with re-

Table 6: Distribution of spelling errors for 3 110 company identifiers in the USPTO patents. The mean of spelling errors per company identifier and the standard deviation σ refer to companies with misspellings. The last row in each table shows the number of patents that are additionally found if the original query q is expanded by A_q^*.

(a) Distribution of spelling errors between the years 2001 and 2010.

		Year									
		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Measure	Number of companies	1 028	1 066	1 115	1 151	1 219	1 261	1 274	1 210	1 224	1 268
	Number of companies with misspellings	67	63	53	65	65	60	65	64	53	60
	Companies with misspellings (%)	6.52	5.91	4.75	5.65	5.33	4.76	5.1	5.29	4.33	4.73
	Mean	2.78	2.35	2.23	2.28	2.18	2.48	2.23	3.0	2.64	2.8
	Standard deviation σ	4.62	3.3	3.63	3.13	2.8	3.55	2.87	6.37	4.71	4.6
	Maximum misspellings per company	24	12	16	12	10	18	12	45	28	22
	Additional number of patents	7.1	7.21	7.43	7.68	7.91	8.48	7.83	8.84	8.92	8.92

(b) Distribution of spelling errors based on the IPC scheme.

		IPC code							
		A	B	C	D	E	F	G	H
Measure	Number of companies	954	1 231	811	277	412	771	1 232	949
	Number of companies with misspellings	59	70	51	7	10	33	83	63
	Companies with misspellings (%)	6.18	5.69	6.29	2.53	2.43	4.28	6.74	6.64
	Mean	3.0	2.49	3.57	1.86	2.8	1.88	3.29	4.05
	Standard deviation σ	5.28	3.65	7.03	1.99	4.22	2.31	5.72	7.13
	Maximum misspellings per company	32	14	40	3	12	6	24	35
	Additional number of patents	9.25	9.67	11.12	4.71	4.6	4.79	8.92	12.84

spect to the Fortune 500 sample (cf. Table 4), where company names that are longer and presumably more difficult to write contain more spelling errors.

In contrast to the uniform distribution of misspellings over the years, the situation with regard to the technological areas is different (cf. Table 6(b)). Most companies are associated with the IPC sections G and B, which both refer to technical domains (cf. Table 6(b), Row 1). The percentage of misspellings in these sections increased compared to the spelling errors grouped by year. A significant difference can be seen for the sections D and E. Here, the number of assigned companies drops below 450 and the percentage of misspellings decreases significantly from about 6% to 2.5%. These findings might support the hypothesis that spelling errors are inserted deliberately in technical domains.

5 Conclusions

While researchers in the patent domain concentrate on retrieval models and algorithms to improve the search performance, the original aspect of our paper is that it points to a different (and orthogonal) research avenue: the analysis of patent inconsistencies. With the analysis of spelling errors in assignee names we made a first yet considerable contribution in this respect; searches with assignee constraints become a more sensible operation. We showed how a special treatment of spelling errors can significantly raise the effectiveness of patent search. The identification of this untapped potential, but also the utilization of machine learning to combine patent features with typography, form our main contributions.

Our current research broadens the application of a patent spelling analysis. In order to identify errors that are introduced deliberately we investigate different types of misspellings (edit distance versus phonological). Finally, we consider the analysis of acquisition histories of companies as promising research direction: since acquired companies often own granted patents, these patents should be considered while searching for the company in question in order to further increase the recall.

Acknowledgements

This work is supported in part by the German Science Foundation under grants STE1019/2-1 and FU205/22-1.

References

Stephen Adams. 2010. The Text, the Full Text and nothing but the Text: Part 1 – Standards for creating Textual Information in Patent Documents and General Search Implications. *World Patent Information*, 32(1):22–29, March.

Mikhail Bilenko and Raymond J. Mooney. 2002. Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases. Technical Report AI 02-296, Artificial Intelligence Laboratory, University of Austin, Texas, USA, Austin, TX, February.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *EMNLP-CoNLL '07: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 858–867. ACL, June.

Qing Chen, Mu Li, and Ming Zhou. 2007. Improving Query Spelling Correction Using Web Search Results. In *EMNLP-CoNLL '07: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 181–189. ACL, June.

Peter Christen. 2006. A Comparison of Personal Name Matching: Techniques and Practical Issues. In *ICDM '06: Workshops Proceedings of the sixth IEEE International Conference on Data Mining*, pages 290–294. IEEE Computer Society, December.

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In Subbarao Kambhampati and Craig A. Knoblock, editors, *IIWeb '03: Proceedings of the IJCAI workshop on Information Integration on the Web*, pages 73–78, August.

Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7(3):171–176.

Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16.

Caspas J. Fall and Christophe Giraud-Carrier. 2005. Searching Trademark Databases for Verbal Similarities. *World Patent Information*, 27(2):135–143.

Matthias Hagen and Benno Stein. 2011. Candidate Document Retrieval for Web-Scale Text Reuse Detection. In *18th International Symposium on String Processing and Information Retrieval (SPIRE 11)*, volume 7024 of *Lecture Notes in Computer Science*, pages 356–367. Springer.

David Hunt, Long Nguyen, and Matthew Rodgers, editors. 2007. *Patent Searching: Tools & Techniques*. Wiley.

Intellevate Inc. 2006. Patent Quality, a blog entry. `http://www.patenthawk.com/blog/2006/01/patent_quality.html`, January.

Hideo Joho, Leif A. Azzopardi, and Wim Vanderbauwhede. 2010. A Survey of Patent Users: An Analysis of Tasks, Behavior, Search Functionality and System Requirements. In *IIix '10: Proceeding of the third symposium on Information Interaction in Context*, pages 13–24, New York, NY, USA. ACM.

Donald E. Knuth. 1997. *The Art of Computer Programming, Volume I: Fundamental Algorithms, 3rd Edition*. Addison-Wesley.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Original in *Doklady Akademii Nauk SSSR* 163(4): 845-848.

Yanen Li, Huizhong Duan, and ChengXiang Zhai. 2011. CloudSpeller: Spelling Correction for Search Queries by Using a Unified Hidden Markov Model with Web-scale Resources. In *Spelling Alteration for Web Search Workshop*, pages 10–14, July.

Patrice Lopez and Laurent Romary. 2010. Experiments with Citation Mining and Key-Term Extraction for Prior Art Search. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *CLEF 2010 LABs and Workshops, Notebook Papers*, September.

Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, editors. 2011. *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*. Springer.

Walid Magdy and Gareth J. F. Jones. 2010. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *CLEF 2010 LABs and Workshops, Notebook Papers*, September.

Walid Magdy and Gareth J.F. Jones. 2011. A Study on Query Expansion Methods for Patent Retrieval. In *PAIR '11: Proceedings of the 4th workshop on Patent information retrieval*, AAAI Workshop on Plan, Activity, and Intent Recognition, pages 19–24, New York, NY, USA. ACM.

Alvaro E. Monge and Charles Elkan. 1997. An Efficient Domain-Independent Algorithm for Detect-

ing Approximately Duplicate Database Records. In *DMKD '09: Proceedings of the 2nd workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 23–29, New York, NY, USA. ACM.

Heiko Müller and Johann-C. Freytag. 2003. Problems, Methods and Challenges in Comprehensive Data Cleansing. Technical Report HUB-IB-164, Humboldt-Universität zu Berlin, Institut für Informatik, Germany.

Felix Naumann and Melanie Herschel. 2010. *An Introduction to Duplicate Detection*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

Yoh Okuno. 2011. Spell Generation based on Edit Distance. In *Spelling Alteration for Web Search Workshop*, pages 25–26, July.

Martin Potthast and Benno Stein. 2008. New Issues in Near-duplicate Detection. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications. Selected papers from the 31th Annual Conference of the German Classification Society (GfKl 07)*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 601–609, Berlin Heidelberg New York. Springer.

Benno Stein and Daniel Curatolo. 2006. Phonetic Spelling and Heuristic Search. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *17th European Conference on Artificial Intelligence (ECAI 06)*, pages 829–830, Amsterdam, Berlin, August. IOS Press.

Benno Stein and Matthias Hagen. 2011. Introducing the User-over-Ranking Hypothesis. In *Advances in Information Retrieval. 33rd European Conference on IR Resarch (ECIR 11)*, volume 6611 of *Lecture Notes in Computer Science*, pages 503–509, Berlin Heidelberg New York, April. Springer.

U.S. Patent & Trademark Office. 2010. Manual of Patent Examining Procedure (MPEP), Eighth Edition, July.

William W. Winkler. 1999. The State of Record Linkage and Current Research Problems. Technical report, Statistical Research Division, U.S. Bureau of the Census.

Xiaobing Xue and Bruce W. Croft. 2009. Automatic Query Generation for Patent Search. In *CIKM '09: Proceeding of the eighteenth ACM conference on Information and Knowledge Management*, pages 2037–2040, New York, NY, USA. ACM.