Semi-automatic Detection of Cross-lingual Marketing Blunders based on Pragmatic Label Propagation in Wiktionary

Christian M. Meyer and **Judith Eckle-Kohler** and **Iryna Gurevych**
Ubiquitous Knowledge Processing (UKP) Lab
and Research Training Group AIPHES
Technische Universität Darmstadt, Germany
https://www.ukp.tu-darmstadt.de

Abstract

We introduce the task of detecting cross-lingual marketing blunders, which occur if a trade name resembles an inappropriate or negatively connotated word in a target language. To this end, we suggest a formal task definition and a semi-automatic method based the propagation of pragmatic labels from Wiktionary across sense-disambiguated translations. Our final tool assists users by providing clues for problematic names in any language, which we simulate in two experiments on detecting previously occurred marketing blunders and identifying relevant clues for established international brands. We conclude the paper with a suggested research roadmap for this new task. To initiate further research, we publish our online demo along with the source code and data at http://uby.ukp.informatik.tu-darmstadt.de/blunder/.

1 Introduction

Large companies increasingly advertise and sell their products in international markets. Developing a marketing campaign for a new country requires tremendous translation efforts in order to bridge language and cultural boundaries. A particular problem often occurs if an established product, brand, or company name is introduced to a new, foreign market without being adapted to local habits and language use. This may yield offensive, embarrassing, or (at best) funny results causing excessive remedial cost and maybe even the withdrawal of a product from the new market. Such a *marketing blunder* can have multiple different reasons. A commercial for a men's fragrance showing a man with his dog failed, for instance, in Islamic countries where dogs are considered unclean. Dalgic and Heijblom (1996) distinguish possible reasons, including political, ethical, and legal issues, different traditions, inappropriate language, etc.

In this work, we focus on *cross-lingual marketing blunders*, which are a result of using inappropriate or negatively connotated expressions or translations for naming a company, brand, or product. One example for this is using the word *mist*, which usually describes fabulous, enigmatic, lightweight, or mystic things in English. A British car manufacturer, for example, chose the word to advertise their *Silver Mist* model. In German, the false friend *Mist* means, however, dung or manure, and it is a frequently used slang expression to describe a futile, cheap, or broken thing, nonsense, or an annoying, tedious situation. This pejorative meaning has caused the car manufacturer to rename its product (Room, 1982; Felser, 2010). A more recent example is the announcement of the 7th edition of a smartphone in Asia. The company's original English slogan "This is 7" has been changed for the Hong Kong market, as the pronunciation of the numeral seven (jyutping: *cat1*) is very similar to a vulgar expression for the male genitals (*cat6*), which would cause funny reactions when combined with "This is" on a product's advertisement poster.[1]

Spotting a marketing blunder can be very time-consuming and expensive for companies, especially if they do not operate local branches in all their target countries. With the emergence of the world wide web, a myriad of start-ups and small companies is struggling with this issue when planning an international online shop. They face two major problems:

(1) The absence of large-scale resources yielding *clues* for potential marketing blunders. Since many blunders are caused by false friends used in colloquial speech, multilingual dictionaries covering the

This work is licensed under a Creative Commons Attribution 4.0 International License.
License details: http://creativecommons.org/licenses/by/4.0/

[1] See for example http://qz.com/777628/ (September 9, 2016)

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2071–2081, Osaka, Japan, December 11-17 2016.

standard language are of limited help. Although there are specialized monolingual slang dictionaries such as the *McGraw-Hill's American Slang Dictionary* (Spears, 2007) for the U.S., it is very challenging to keep these dictionaries up-to-date and to provide them for a large number of languages.

(2) The absence of tools that assist the process of identifying the *relevant clues* from these resources. Obviously, not every word that exists in a target language is problematic for marketing a product. The English word *fog* is, for instance, a false friend of the Hungarian *fog* (English: *tooth*). Neither meaning has a negative connotation per se that would impede the use of *fog* in a successful marketing campaign within those countries. A tool assisting the detection of marketing blunders thus needs to separate relevant clues from irrelevant ones in order to reduce the manual effort.

In the present paper, we propose a novel method and tool for assisting copywriters and sales promoters with the detection of cross-lingual marketing blunders. Our method is primarily based on disambiguated translations and pragmatic labels extracted from Wiktionary (`http://www.wiktionary.org`), for which we create a large inter-lingual index and a retrieval process. We evaluate our approach in two experiments: (1) detecting previously occurred marketing blunders and (2) finding evidence for potential blunders in established brand names. The detection of cross-lingual marketing blunders is a new task in natural language processing and, to the best of our knowledge, there are yet no existing tools that assist copywriters in avoiding such blunders. This is why we aim at introducing a formal task definition, a first, freely available dataset, and a novel knowledge-based method to initiate further research in this direction. Based on our results and error analysis, we lay out a research agenda for this new task. Apart from detecting marketing blunders in trade names, we believe that this research strand is enabling for many other tasks, including the identification of problematic product slogans, acronyms (e.g., of scientific proposals), and names of persons, institutions, and projects that should not be misinterpreted in foreign languages.

2 Related Work

Marketing blunders are yet mostly discussed in management and marketing research. Ricks (2006) reports a large number of previously occurred blunders in international business, including a separate chapter on product and company names. Knight (1995) and Dalgic and Heijblom (1996) discuss a few number of cases in detail. These works aim at finding new management strategies for inter-cultural marketing (cf. Jallat and Kimmel, 2002), rather than providing actual assistance and tools for copywriters.

In another strand of research, linguists have studied the properties of the language used in advertising, including teasers, slogans, and names. Cook (1992) and Janich (2013) give comprehensive introductions to the linguistic analysis of marketing language. While these works are mostly concerned with the question of how positive connotations and rhetorical figures enhance the value of a product, they also touch on the issues of cross-cultural and cross-lingual marketing communication. However, none of these works describes specific properties or methods to detect potential naming blunders. In addition to that, there are dictionaries on slang and pejorative expressions like the ones by Spears (2007) or Küpper (1984), as well as specialized dictionaries on trade names, such as Room (1982). Slang dictionaries are limited in their up-to-dateness (as indicated by the old publication years), word coverage, and range of available languages. Specialized name dictionaries are generally of little use for this task, as it is often the essence of a marketing campaign to create new, previously unused product or brand names.

In natural language processing, there are previous works which address the automatic generation and retrieval of slogans and creative names. Veale (2011) presents a search engine for creative text retrieval, which assists copywriters to find metaphors and unusual word combinations. Özbal and Strapparava (2012) describe an automatic approach to generate neologisms that can be used as product names or slogans. Both works are focused on the English language and on the identification or generation of *good* names, whereas our work aims at the detection of *problematic* names without focusing on a particular language or target market. Our task is similar to automatically distinguishing cognates and false friends. Inkpen et al. (2005) use orthographic similarity metrics for this task including Soundex, which we also propose for our method. Follow-up works by Mitkov et al. (2007), Gomes and Lopes (2011), Beinborn et al. (2013), and Ciobanu and Dinu (2014) propose different edit distance, machine translation, and semantic similarity methods for identifying cognates. While false friends play a crucial role for the detection of marketing blunders, the existing approaches cannot be used directly, because they are specific for a certain language pair and do not make any assumptions on the (negative) connotations of a trade name. Kondrak and Dorr (2004) adapt orthographic and phonetic similarity methods to find confusable drug names. Unlike our tool which retrieves pragmatically marked sense descriptions, their work is limited to identifying similar forms.

The recognition of words and phrases associated with opinions and emotions is the goal of sentiment analysis. There have been multiple attempts to construct large sentiment lexicons, such as SentiWordNet (Esuli and Sebastiani, 2006). Banea et al. (2008) propose a bootstrapping approach to induce such lexicons from a small, manually defined seed list. Our approach is similar, since we propagate pragmatic information based on lexical relations. However, we do not require a seed list and we consider relations across many languages. While monolingual resources are not suitable for our task at all, the existing multilingual sentiment lexicons are severely limited in size. The NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013) is among the largest and available in about 40 languages, but since it has been automatically translated, it does not distinguish word senses and lacks slang and dialects. In recent work, Vo and Zhang (2016) propose a neural network architecture to build sentiment lexicons relying on emoticons as distant supervision signals. Although they currently publish only English and Arabic lexicons, such (almost) unsupervised methods are promising for building multilingual lexicons.

3 Task Formalization

Let T denote a product, brand, or company name. Our goal is to develop a method $\mathcal{M}$ retrieving a set of clues $C = \mathcal{M}(T)$ for a given T, which can be used by copywriters to decide whether T should be accepted or rejected as a name. We consider each clue $c = (w, \ell, d) \in C$ as a tuple of a word form w of language ℓ and a textual description d, which paraphrases the (potentially problematic) meaning of w in ℓ. While T is not specific to any language, a copywriter is typically only fluent in a few number of languages, which is why d must be in a language spoken by the end user (hereafter *output language*). For the example $T =$ "Silver Mist", a method could return the following clues:

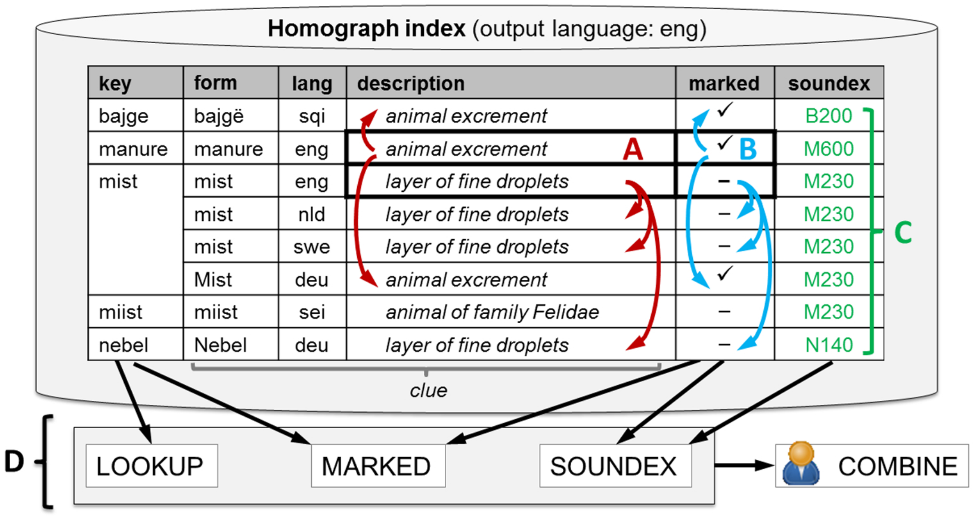
#	form w	language ℓ	description d
1	Silber	German (deu)	A shiny gray color.
2	mist	English (eng)	A layer of fine droplets or particles.
3	Mist	German (deu)	Manure; animal excrement.
4	miist	Seri (sei)	An animal of the family Felidae.
5	miste	Danish (dan)	To lose something.
6	silver mine	English (eng)	A mine for silver ore.

The first clue refers to the German translation of *silver*, which has the similar word form *Silber*. The second and third clues address the word *mist* with its meanings in English and German. The fourth and fifth clues refer to similar word forms of *mist* in Danish and Seri (an isolated language spoken in Mexico). The last clue addresses a multi-word expression which has a similar form to the entire name T. When deciding if T should be rejected as a name, only the third and fifth clues are helpful, since the specified meanings are negatively connotated and thus do not enhance the value of a product with such a name. We say a clue c is *relevant* for T if it should be considered in the decision of accepting or rejecting the name. Note that there might be good reasons for a copywriter to ignore the fact that *Mist* has a negatively connotated meaning in German. In section 5, we discuss such cases.

The primary objective of a method $\mathcal{M}$ is to return relevant clues for as many names as possible and thus obtain a high recall. While this is obviously important for maximizing the usefulness of the method, a secondary objective is to retrieve *only* relevant clues and thus obtain a high precision. The rationale behind this is to minimize the amount of information that needs to be manually checked by humans.

4 Proposed Method

Our solution is based on the following considerations:



key	form	lang	description	marked	soundex
bajge	bajgë	sqi	*animal excrement*	✓	B200
manure	manure	eng	*animal excrement*	✓	M600
mist	mist	eng	*layer of fine droplets*	–	M230
	mist	nld	*layer of fine droplets*	–	M230
	mist	swe	*layer of fine droplets*	–	M230
	Mist	deu	*animal excrement*	✓	M230
miist	miist	sei	*animal of family Felidae*	–	M230
nebel	Nebel	deu	*layer of fine droplets*	–	N140

Figure 1: Running example of our method showing sense definition (A) and pragmatic label propagation (B), Soundex representation (C), and the practical usage scenario based on different query methods (D)

(1) The input name T can be of any language. Approaches using *monolingual* corpora or knowledge bases are therefore of little help or yield a tool that can only be used for a single target market. Instead, we aim at covering many languages and markets.

(2) The description of a clue should be limited to a few predefined output languages understood by the copywriters using the system.

(3) A solution must deal with non-standard language varieties, such as slang and dialects, to detect vulgarities, negative connotations, etc.

(4) Crowdsourcing platforms might prove helpful, but do not return instant feedback, require to divulge T before its official announcement, and are highly biased towards certain languages (cf. Pavlick et al., 2014).

(5) We lack training data for the marketing blunder task, which is why we do not consider data-driven or machine-learning approaches yet. Large annotated multilingual datasets including non-standard language will be necessary to learn a generalizing model (cf. section 7).

We propose using data from Wiktionary, which is particularly suitable for this task, since it contains lexicon entries and translations in many languages and a broad diversity of technical domains, colloquial language, slang, and dialects (cf. Meyer and Gurevych, 2012). As the users of our tool, we assume copywriters speaking English and German, which is a realistic example for central European marketing agencies. Our method is, however, not limited to these two output languages.

In the remaining section, we describe our family of methods $\mathcal{M}(T)$, which first segment the input name T into a sequence of k tokens $T = t_1 t_2 \ldots t_k$. The methods then create the set Q of all token n-grams in T and retrieve clues for each $q \in Q$ from a huge inter-lingual index. We create this so-called *homograph index* by extracting and propagating sense definitions, translations, and pragmatic labels from Wiktionary.[2] The four steps of this approach are explained below and summarized in figure 1. We use the running example T = "Silver Mist", for which our methods retrieve clues from the homograph index for each token n-gram $q \in Q = \{$*Silver*, *Mist*, *Silver␣Mist*$\}$.

Step A: Propagating sense definitions. The example of the *Silver Mist* car suggests that a large share of cross-lingual marketing blunders is due to false friends (i.e., two words with the same pronunciation or written form, but different meanings). This is why we first create a *homograph index* of words sharing the same word form. Each index entry consists of a normalized word form (the *key*) that points to one or multiple clues (i.e., triples of word form w, language ℓ, and textual description d).

Initially, the homograph index contains only clues extracted from all Wiktionary word senses of the

[2] We use the DKPro JWKTL software to extract this information: https://dkpro.github.io/dkpro-jwktl/

output languages (e.g., English and German). We create the key by converting the lemma of the Wiktionary word sense to lower case and removing special characters and diacritics. As the textual description d of the corresponding clues, we use the senses' definitions. In a subsequent step, we extend the homograph index by adding all translations of these word senses. We apply the same normalization technique for the translated word form to obtain the key. The key *han* thus points to clues in eleven languages, including the lemma forms *Han* (e.g., English), *han* (Turkish), *hän* (Finnish), *hăn* (Frisian), *hån* (Swedish), and *hẳn* (Vietnamese). This multilingual homograph index already enables queries in any language for which translations are encoded in Wiktionary. In order to determine the clues for the translated word forms, we propagate the sense definition from the output language to the translation language. This is possible, because each Wiktionary translation is associated with a specific word sense. This way, copywriters can decide to accept or reject a name using a proper sense description instead of a bare word translation, which is especially necessary for polysemous words: Showing only the English word *arm* as a translation of the Polish *broń* would not be helpful, as it remains unclear if the potentially unwanted weapon sense or the unproblematic body part sense of *arm* is meant. In figure 1 (A), we show the homograph index creation for the two English output language lemmas *manure* and *mist*. We add the translations of the corresponding Wiktionary word senses as new index entries, and we propagate the sense description from the thick-framed cells to the foreign language entries.

Our final homograph index consists of 1.3 million normalized word forms referring to about 3.0 million clues covering 2,022 languages.[3] Using this index, we can define our first method for retrieving clues, which we call LOOKUP. Given the input name T and the set of all token n-grams Q, the LOOKUP method normalizes each $q \in Q$ using the same technique as for index keys and then looks up every normalized q in our index. For $T =$ "Silver Mist", LOOKUP returns 48 clues from six languages.

Step B: Propagating pragmatic labels. An important goal of our approach is separating relevant from irrelevant clues. The Dutch and Swedish forms of *mist* are, for instance, cognates of the English word form and thus carry the same, unproblematic meaning. Likewise, false friends without any negative connotation, such as the English and Hungarian *fog* discussed above, yield irrelevant clues that should be ignored. In a dictionary, the corresponding entries for these words are usually *unmarked* – i.e., they are not associated with a particular language variety, but considered standard language. As opposed to that, the German *Mist* is marked as "umgangssprachlich" ("slang") and as "verärgerte Äußerung" ("annoyed utterance") in Wiktionary, which are good indicators to avoid using *Mist* in a product name. We call these markings *pragmatic labels* (Wiegand et al., 2010). For our purposes, we are interested in sociological labels (the diastratic variety) that mark jargon used by a certain culture, social group, or social class (e.g., *army slang*, *argot*, *children's language*), register and style labels (the diaphasic variety) that mark word senses used in certain communicative situations (e.g., *colloquial*, *informal*, *slang*), and evaluative labels (the diaevaluative variety) that mark offensive words and words with a certain connotation (e.g., *pejorative*, *rude*, *derogatory*). Note that this goes beyond sentiment lexicons, which typically focus on the diaevaluative variety. In Wiktionary, pragmatic labels are specified at the beginning of a sense definition, usually enclosed in parentheses, typed in italics, or separated by a colon. We extract all the labels used for the word senses of the output languages. Of the 2,440 distinct labels used at least three times, we manually select 245 labels that belong to one of the three label categories and we enrich our homograph index by storing whether a word sense is marked by one of these labels. In a subsequent step, we propagate the labels from the output languages to the other languages by following the translation links. The underlying assumption is that a pragmatic label of a word sense in one language is conserved in another language, given that the translation is correct. Figure 1 (B) shows an example for English: The output language word sense of *manure* is marked as slang. We propagate this marking to the equivalent German word sense *Mist* and the Albanian (sqi) word sense *bajgë*. About 63,000 clues of the homograph index are marked by at least one of the 245 pragmatic labels. Based on this, we define a second method called MARKED, which looks up each normalized token n-gram of Q in the homograph index (equivalent to LOOKUP), but returns only clues marked by a pragmatic label.

[3] But note the long tail: 1,015 languages have only one clue.

Step C: Similar word form representations. Marketing blunders are of course not limited to forms with the exact same written form. Choosing the word forms *misd*, *misth*, or *miist* could, for instance, cause similar reactions in German-speaking markets as their pronunciation is highly similar to *Mist*. This is why we propose a third method SOUNDEX, which queries the homograph index for marked word forms starting with the same three letters and having the same *Soundex* representation (Russell, 1918) as the queried token n-gram (but not being identical to it). Soundex is an algorithm that returns a pseudo-phonetic representation of a given English word, which consists of the word's initial letter followed by at least three digits denoting groups of similar consonants. The main idea of the algorithm is that two words with similar pronunciations return similar Soundex representations: The Soundex representation of both *mist* and *miist* is M230, but M62352 for *marketing*. Although the Soundex algorithm is designed for the English language, we apply it to any token n-gram and leave the development of a language-independent pseudo-phonetic model to future research. To allow for faster queries, we precompute the Soundex representations for all clues of our homograph index as shown in figure 1 (C). When using the SOUNDEX method, we can then create a Soundex representation for each token n-gram $q \in Q$ and retrieve all marked clues that have the same Soundex representation in a simple homograph index lookup.

Step D: Practical, semi-automatic usage scenario. We combine the three methods by first querying the homograph index using MARKED. If this search does not yield relevant clues, we query the index using SOUNDEX and, analogously, we query the index using the LOOKUP method if there are still no relevant clues. We thus define the method COMBINE as the sequential application of MARKED, SOUNDEX and LOOKUP, see figure 1 (D). The rationale behind this is to simulate a practical usage situation of our approach: Since only a fraction of the entries are marked by pragmatic labels, we first present those to a user (MARKED). If she or he finds evidence for a marketing blunder, no further lookup is required for the given name. Otherwise, the user can check for marked forms with a similar form representation (SOUNDEX) and only turn towards reading all entries (LOOKUP) if there are still no relevant clues.

5 Evaluation

To evaluate our approach, we conduct two experiments: (1) We measure the performance of our four methods using a newly created dataset of previously occurred cross-lingual marketing blunders. (2) We apply our method to a large dataset of international brand names to check for potential blunders.

Marketing blunder dataset. As a novel evaluation dataset for this task, we extract the marketing blunder examples discussed by Ricks (2006, § 3) and provided on the homepage of the British consultancy *Commisceo Global*.[4] We omit examples that are not related to a name or whose name or translation is not explicitly provided. Ricks notes, for instance, that a U.S. food manufacturer wrongly translated the name of their mascot "Jolly Green Giant" into Arabic as "intimidating green ogre", but does not provide the exact Arabic form, which would be required to properly simulate the tool-assisted detection of this blunder. For each blunder in this dataset, we store the problematic name T, a remark on the vendor or type of product, and a short textual explanation of the blunder. In addition to that, we manually group the blunders into the following four categories:

- *vulgar:* names containing vulgar, rude, or offensive expressions,
- *sexual:* names with sexual innuendos,
- *negative:* names with negative connotations or suggesting negative properties,
- *intent:* names containing an expression with a different, unrelated meaning in a certain language causing astonishment and distraction among potential customers.

The fruit drink *Pavian* is an example for the *intent* group since *Pavian* means *baboon* in German, which caused distraction among customers, although the word is not negatively connotated. Our initial dataset consists of 44 cross-lingual marketing blunders, which we make publicly available for other researchers.

Experiment 1. We apply our four methods to each problematic name of this novel marketing blunder dataset. In total, our methods returned 1,494 clues. In order to judge a clue relevant or irrelevant, we

[4] http://www.commisceo-global.com/blog/cross-cultural-marketing-blunders (accessed: 2016-05-02)

	MARKED	SOUNDEX	LOOKUP	COMBINE
Detected marketing blunders:	18 / 44	18 / 44	28 / 44	34 / 44
intent	0 / 3	0 / 3	3 / 3	3 / 3
negative	3 / 15	8 / 15	10 / 15	14 / 15
sexual	7 / 14	4 / 14	7 / 14	8 / 14
vulgar	8 / 12	6 / 12	8 / 12	9 / 12
Retrieved clues:	105 / 151	85 / 247	341 / 1202	229 / 517
Precision P:	.70	.34	.28	.44
Recall R:	.41	.41	.64	.77
F_1 score:	.52	.37	.39	.56
F_2 score:	.45	.39	.51	.67

Table 1: Evaluation results for our marketing blunder dataset

ask two human raters to annotate this set of retrieved clues. The raters agree on 95 % of the judgments yielding an inter-rater agreement of $\kappa = .87$ (using Cohen's kappa). Based on this agreement, we consider the annotations reliable (cf. Artstein and Poesio, 2008). For obtaining a gold standard, we ask an additional adjudicator to decide on the 76 ties.

Table 1 summarizes the evaluation results. We report the number of detected blunders over the total number of blunders both for the whole dataset and separately for each blunder category. The table additionally provides the total number of relevant and retrieved clues as well as the precision, recall, F_1 and F_2 scores. We define the precision P as the ratio of relevant clues to the total number of retrieved clues (i.e., a method is more precise if it returns more relevant clues) and recall R as the proportion of detected marketing blunders in the dataset (i.e., a method has a higher recall if it is able to detect more marketing blunders). The F_1 and F_2 scores follow the standard definitions of being the (weighted) harmonic mean between precision and recall. In accordance with our task's primary objective (see section 3), the F_2 score prefers high recall over high precision.

The basic LOOKUP method yields a recall of .64 indicating that our homograph index is able to effectively detect cross-lingual marketing blunders. As the low precision indicates, there are, however, a large number of irrelevant clues that are retrieved by this simple index lookup. As opposed to that, we find a high precision for the MARKED method, as the index entries marked with pragmatic labels yield relevant clues in over 70 % of the cases. The MARKED method is, however, not suitable to detect marketing blunders of the categories *intent* and *negative*, which causes a low recall. It should be noted that it is not surprising to find the recall of MARKED lower than that of LOOKUP, because the former returns a subset of the latter. This is different for SOUNDEX, which facilitates the detection of marketing blunders that remain unseen by the other methods. We find that our semi-automatic usage simulation COMBINE yields the most reasonable trade-off between precision and recall, since it achieves the highest recall of the three methods and a higher precision than LOOKUP and SOUNDEX. With this method, we are able to detect 34 of the 44 cross-lingual marketing blunders ($R = .77$) and we achieve the highest F_1 and F_2 scores. For the corresponding marketing campaigns, the copywriters would have to examine a total of 517 clues (on average 12 per blunder); 229 of them are relevant (on average 5 per blunder; $P = .44$).

Experiment 2. In our second experiment, we aim at finding potential marketing blunders in existing names on a larger scale. To this end, we use all 998 brand names of the BrandPitt corpus (Özbal et al., 2012) and retrieve clues using our tool. It is important to note that this corpus mostly contains top-tier international brands, such as *Pizza Hut*, *Jaguar*, and *IKEA*, whose names are established for many years and thought over by leading marketing agencies. Initially, we therefore did not expect to find many relevant clues. Our tool returns a total of 756 clues using MARKED, 3,549 using SOUNDEX, and 17,270 using LOOKUP. Given the high number of brand names and clues, we focus on the clues returned by MARKED (i.e., the first step of our simulation) and ask two human raters to judge them relevant or irrelevant for deciding whether or not to use a name for a particular region in the world. The raters find

154 (rater 1) and 192 (rater 2) of the 756 clues to be relevant. This corresponds to returning relevant clues for 70 (rater 1) and 88 (rater 2) of the 215 names for which MARKED returned at least one clue. The raters agree in 90 % of the cases ($\kappa = .72$).

Finding this many relevant clues is surprising, which is why we carefully checked the annotated data. We indeed find very helpful clues, which – in our opinion – should at least be known to the corresponding marketing agencies. The name of the animation studio *Pixar* means, for example, to urinate in the Catalan language. The name of the Norwegian confectionery *NERO* and the German software producer *Nero* means *brainiac* in Finnish, which is marked derogatory in some contexts, and the related form *ñero* is a rough equivalent of *thug* or *gangsta* in Colombia. The name of the Russian car manufacturer *Lada* has a related form *låda* in Swedish meaning *box*, including a pejorative meaning for unattractive houses and cars. Though being differently pronounced, the relationship could still be problematic in written communication (e.g., a poster with barely visible ring diacritic). In the Darfur and Chad language Fur, *martìn* is a vulgar form for the buttocks, which the raters consider relevant for the *Aston Martin* car brand. The popular coffee bar name *Thanks a Latte* might be less suitable for the German market, where *Latte* is a colloquial word for erected penis. The name of the *Coco Pops* cereals might prove problematic in French markets, where *coco* means cocaine. Though being relevant, many of these clues are of course not problematic, since the brand names are already well-established. However, we consider our tool helpful for new names, which lack this brand strength. Another important finding from this experiment is that there are often relevant clues for which a more salient word sense exists that prevents misinterpretation. The related forms *cocó* and *cocô* mean, for instance, *shit* in Portugal and Brazil, respectively, which we consider highly relevant for the *Coco Pops* example. Since there is, however, also the frequent form *coco* in both languages meaning coconut, it is unlikely that the product name is misinterpreted.

6 Discussion and Error Analysis

MARKED works well for detecting the blunder categories *vulgar* and *sexual*, whereas LOOKUP predominantly retrieves clues for the *intent* and *negative* categories. Since we designed SOUNDEX to only return marked entries, it is likewise less suitable for *negative* and *intent*. For *negative* blunders, additional knowledge from sentiment lexicons could be helpful to recognize relevant clues containing a negative connotation. In order to do so, we require either language-independent sentiment analysis tools or a sense-disambiguated notion of sentiment for Wiktionary word senses, which can be used for propagating sentiment information across languages. Existing lexicons, such as the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013), provide a good starting point.

Blunders of the *intent* category are much harder and most likely require copywriters to read all clues returned by LOOKUP. Semantic relatedness measures might prove useful for identifying false friends with highly different meanings. Especially in the BrandPitt dataset, we note, however, that there are some names using pragmatically marked or ambiguous words intentionally. The *Get Lost Magazine*, for instance, includes the English phrase *get lost* which raises negative associations of rudely being asked to leave. Since the magazine is about adventure traveling, the copywriters use this name on purpose to create an interesting name with multiple interpretations. This illustrates why we consider it important to model marketing blunder detection as a semi-automatic task leaving the final decision to humans.

None of our current methods is able to detect blunders whose text contains a problematic word as a substring. The product name *FARTFULL*, for instance, needs to be split into *fart* and *full*, before an index lookup can yield relevant clues. The large number of substring combinations would, however, yield a huge number of irrelevant clues if all combinations are queried. The English words *fartherer* or *penny-farthing* contain, for instance, the substring *fart*, but do not lead to a vulgar interpretation right away. A similar problem occurs for inflected word forms (e.g., *Vicks*). Relying on automatic lemmatization or morphological analysis is problematic, since such tools are usually language-specific. For our task, they would need to cover essentially any language. Trimming suffixes of different lengths might be a solution, but would not work for highly agglutinative languages.

The use of Soundex representations for identifying similar forms works well in some cases, for example for finding the Finnish *hullu* (IPA: [ˈhulːu]) meaning insane, which has a similar pronunciation as the American streaming service *Hulu* ([huˈlu]). For the Japanese *creap*, SOUNDEX retrieves the English *creep* (annoying person) and the French *crevé* (extremely fatigued). While the former is relevant for detecting this blunder, the latter might have a somewhat similar pronunciation in English (e.g., [kɹiːv]), but definitely not in French ([kʁə.ve]). The same is true for *Lada* and the Swedish *Låda* ([ˈloː.da]). Such problems are due to Soundex being designed for English. Since Wiktionary also contains pronunciation information, a future method could lookup index entries with similar or equal IPA representation, given that they are available at a large-scale and for a large number of languages. Recently, Deri and Knight (2016) introduced a new grapheme-to-phoneme model for almost any language, which we consider highly relevant for indexing language-independent pseudo-phonetic representations.

Ricks (2006) also notes the English form *crap* as a blunder cause for *creap*, since it has a similar spelling. None of our methods, however, returns *crap* as a clue for this name. An obvious solution would be the use of string similarity metrics, such as Levenshtein's edit distance. We indeed tried this metric, but found that it returns a huge number of irrelevant clues. We therefore suggest to develop a modified edit distance metric for this task, which, for example, puts more weight on editing letters with similar shape. Finally, our tool cannot retrieve clues for potential marketing blunders across different scripts. For the *Bardak* machines, a method would have to transliterate the Latin spelling to the Cyrillic бардак in order to find the problematic meaning of a whorehouse. Future work should incorporate state-of-the-art transliteration systems for as many language pairs as possible.

7 Future Research Demands and Dissemination

Along with this paper, we publish our newly compiled marketing blunder dataset, homograph index, and annotations. In addition to that, we provide a web interface which implements the three steps of our COMBINE method. It allows retrieving clues for arbitrary names. Besides product, brand, and company names, our tool can retrieve clues for acronyms (e.g., of proposals) and person or organization names, which might cause misinterpretations in foreign languages.[5]

With these materials, our Wiktionary-based method and the detailed error analysis, we lay the foundation for the new natural language processing task of detecting marketing blunders. This task raises a number of important research challenges:

- Finding good names is a creative process, for which copywriters intentionally deviate from known patterns. This makes it hard to model the overall detection task with standard pattern recognition algorithms.
- Our task formulation is based on the notion of clues, which explain *why* a name is considered problematic. As opposed to that, many current tasks use a classification setup (e.g., a binary classification into *problematic* and *unproblematic*), which only indicates *if* there is a problem.
- Evaluating a name is highly subjective, as there might be intended ambiguity, jokes and language games, varying association and brand strength, etc. We therefore introduce marketing blunder detection as a semi-automatic task, which are typically very difficult to evaluate.
- Detecting marketing blunders makes most sense if *all* languages are considered, which is especially challenging for poorly documented languages. The necessity to limit the number of output languages raises another interesting challenge of separating object and meta language.

There is a large variety of future projects around this task. First and foremost, we require larger datasets for developing and evaluating our methods. While Wiktionary is continually updated by its community, future versions will yield improved coverage. For supporting additional output languages beyond English and German, it is necessary to scrape other language versions and associate the pragmatic label system with the existing index in a one-time effort. Further knowledge bases and corpora of colloquial language (e.g., from Twitter) may prove useful for background knowledge. For evaluation data, it will be interesting to cooperate with marketing researchers and copywriters. Previously occurred marketing blunders might be collected in a crowdsourcing effort. The second important strand of research will be better blunder detection methods for retrieving relevant clues. The most important issues to solve are

[5] https://github.com/UKPLab/coling2016-marketing-blunders

the intelligent segmentation of names containing problematic words, the automatic transliteration and phonetic representation of the index entries, and the integration of multilingual sentiment lexicons and analysis methods.

8 Conclusion

In this paper, we introduced the task of detecting cross-lingual marketing blunders. In addition to a formal task definition, we proposed a knowledge-based method, which propagates pragmatic labels to translated word senses from Wiktionary. Our final tool assists copywriters who design new names and accept or reject a suggested name based on a number of clues returned by the automatic method. We evaluated our work in two experiments. On a newly created dataset of previously occurred marketing blunders, we were able to detect 78 % of the problematic names. Our second experiment identified between 150 and 200 relevant clues for a large collection of top-tier international brand names.

We find that Wiktionary is well-suited for this task, as it contains many languages and a large number of pragmatic labels allowing us to process non-standard language varieties, such as slang. These language varieties are often absent from newswire corpora and expert-build dictionaries and therefore remain underresearched in our community.

We put a special focus on the error analysis and learned that follow-up work needs to find better tools and methods for computing language- and script-agnostic orthographic and phonetic similarity, for interpreting negative connotations that appear as substrings, and for language-independent sentiment and morphological analysis. To establish the new marketing blunder detection task, we finally discussed its main challenges and demands in order to suggest a research agenda to the scientific community. In future work, it will be especially interesting to cooperate with marketing agencies, in order to study how copywriters use a blunder detection tool for yet unknown product, brand, or company names.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant № I/82806, by the German Research Foundation under grant № GU 798/17-1, and by the German Institute for Educational Research (DIPF). We would like to thank the anonymous reviewers for their helpful comments.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *Proceedings of the Sixth International Language Resources and Evaluation*, pages 2764–2767, Marrakech, Morocco.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate Production using Character-based Machine Translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891, Nagoya, Japan.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Automatic Detection of Cognates Using Orthographic Alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 99–105, Baltimore, MD, USA.

Guy Cook. 1992. *The Discourse of Advertising*. The INTERFACE series. London: Routledge.

Tevfik Dalgic and Ruud Heijblom. 1996. International Marketing Blunders Revisited: Some Lessons for Managers. *Journal of International Marketing*, 4(1):81–91.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-Phoneme Models for (Almost) Any Language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 417–422, Genoa, Italy.

Georg Felser. 2010. Intercultural Marketing. In Alexander Thomas, Eva-Ulrike Kinast, and Sylvia Schroll-Machl, editors, *Handbook of Intercultural Communication and Cooperation*, chapter 1.3, pages 228–242. Göttingen: Vandenhoeck & Ruprecht.

Luís Gomes and José Gabriel Pereira Lopes. 2011. Measuring Spelling Similarity for Cognate Identification. In Luis Antunes and H. Sofia Pinto, editors, *Progress in Artificial Intelligence: Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, volume 7026 of *Lecture Notes in Computer Science*, pages 624–633. Berlin/Heidelberg: Springer.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257, Borovets, Bulgaria.

Frédéric Jallat and Allan J. Kimmel. 2002. Marketing in culturally diverse environments: The case of Western Europe. *Business Horizons*, 45(4):30–36.

Nina Janich. 2013. *Werbesprache: Ein Arbeitsbuch*. Tübungen: Narr, 6th edition.

Gary A. Knight. 1995. Educator Insights: International Marketing Blunders by American Firms in Japan—Some Lessons for Management. *Journal of International Marketing*, 3(4):107–129.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of Confusable Drug Names: A New Approach and Evaluation Methodology. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 952–958, Geneva, Switzerland.

Heinz Küpper. 1984. *Illustriertes Lexikon der deutschen Umgangssprache*. Stuttgart: Klett.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford: Oxford University Press.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Gözde Özbal and Carlo Strapparava. 2012. A Computational Approach to the Automation of Creative Naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 703–711, Jeju Island, Korea.

Gözde Özbal, Carlo Strapparava, and Marco Guerini. 2012. Brand Pitt: A Corpus to Explore the Art of Naming. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1822–1828, Istanbul, Turkey.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

David A. Ricks. 2006. *Blunders in International Business*. Malden: Blackwell Publishing.

Adrian Room. 1982. *Dictionary of Trade Name Origins*. London: Routledge & Kegan Paul.

Robert C. Russell. 1918. *Index*. United States Patent 1,261,167, filed October 25, 1917, published April 2, 1918.

Richard A. Spears. 2007. *McGraw-Hill's American Slang Dictionary*. Chicago: McGraw-Hill, 2nd edition.

Tony Veale. 2011. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 278–287, Portland, OR, USA.

Duy Tin Vo and Yue Zhang. 2016. Don't Count, Predict! An Automatic Approach to Learning Sentiment Lexicons for Short Text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 219–224, Berlin, Germany.

Herbert Ernst Wiegand, Michael Beißwenger, Rufus H. Gouws, Matthias Kammerer, Angelika Storrer, and Werner Wolski, editors. 2010. *Wörterbuch zur Lexikographie und Wörterbuchforschung / Dictionary of Lexicography and Dictionary Research*, volume 1 (Systematische Einführung / Systematic Introduction, A–C). Berlin/New York: de Gruyter.