# Generating titles for millions of browse pages on an e-Commerce site

**Prashant Mathur, Nicola Ueffing and Gregor Leusch**
Machine Translation Science Lab,
eBay Inc.
Kasernenstrasse 25
Aachen, Germany

## Abstract

We present three approaches to generate titles for browse pages in five different languages, namely English, German, French, Italian and Spanish. These browse pages are structured search pages in an e-commerce domain. We first present a rule-based approach to generate these browse page titles. In addition, we also present a hybrid approach which uses a phrase-based statistical machine translation engine on top of the rule-based system to assemble the best title. For the two languages English and German, we have access to a large amount of rule-based generated and human-curated titles. For these languages, we present an automatic post-editing approach which learns how to post-edit the rule-based titles into curated titles.

## 1 Introduction

Natural language generation has a broad range of applications, from question answering systems to story generation, summarization etc. In this paper, we target a particular use case common to many e-Commerce websites. In e-Commerce sites, multiple items can be grouped on a common page called *browse page*. Each browse page contains an overview of various items which share some characteristics. However, the items do not necessarily share all characteristics. The characteristics can be expressed as slot/value pairs. For example, we can have a browse page for all items which share the characteristics (*Watch Type: wrist watch*) and (*Band: stainless steel*). These watches can have additional characteristics such as (*Feature: day clock*) or (*Feature: date indicator*) or (*Feature: chronograph*), which can differ between all items grouped on one browse page. Different combinations of characteristics bijectively correspond to different browse pages, and consequently to different browse page titles. To show customers which items are grouped on a browse page, we need a human-readable description of the content of that particular page. A lack of such description negatively impacts the user experience.

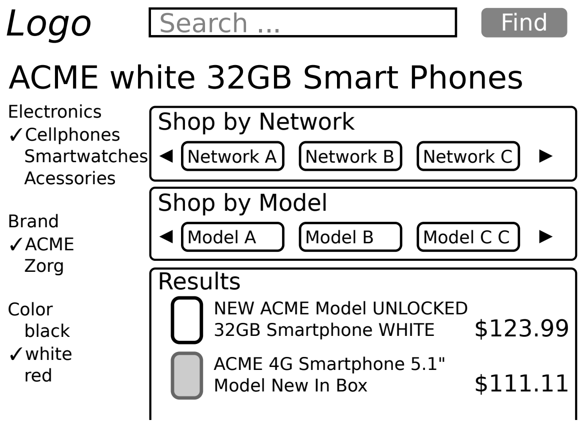
Logo
Search ...
Find
ACME white 32GB Smart Phones
Electronics
✓Cellphones
Smartwatches
Acessories
Brand
✓ACME
Zorg
Color
black
✓white
red
Shop by Network
◄ Network A Network B Network C ►
Shop by Model
◄ Model A Model B Model C C ►
Results
NEW ACME Model UNLOCKED 32GB Smartphone WHITE $123.99
ACME 4G Smartphone 5.1" Model New In Box $111.11

**Figure 1:** Example of a browse page

Figure 1 shows an example of a browse page with a title, with navigation elements leading to related browse pages as well as the individual items listed in this page. The corresponding meta-data for the browse page is shown in Table 1. Note that in our problem definition slot names are already given; the task is to generate a title for a set of slots. Moreover, we do not perform any selection on the slots, all the slots needs to be realized in order to have a unique browse page title.

| Slot Name | Value |
|---|---|
| Category | *Cell Phones & Smart Phones* |
| Brand | *ACME* |
| Color | *white* |
| Storage Capacity | *32GB* |

**Table 1:** The underlying meta-data for Figure 1

We have access to a few thousand human-created titles (curated titles) for English and German. To generate these titles, humans annotators were given a set of slot name-value pairs and were asked to generate a title for that within a set of guidelines, the important one being that all slots need to be realized in the title.

Large e-Commerce sites can easily have tens of millions of such browse pages in many different languages. The problem gets more complex as each browse page can have one to six slots to be realized. On top of that, the number of unique slot-value pairs are in the order of hundreds of thousand. All these factors render the task of human creation of all the browse page titles infeasible. In this paper, we propose several strategies to generate these human-readable titles automatically for any possible browse page, aiming for a high average title quality.

These different strategies address the existence (or non-existence) of human-curated titles across the different languages we are dealing with. In section 3, we present a rule-based system which can be used if there are no human-curated titles at all. We show that title quality can be improved by combining this rule-based system with a machine translation system, exploring monolingual data, as described in section 4. If there are human-curated titles available, we can use automatic post-editing on top of the rule-based system, as described in section 5, to achieve even higher quality titles.

## 2 Previous works

Generating titles for pages containing structured data, for example in e-Commerce, is a frequent problem. Dale et al. (1998) describe the problem of generating natural language titles and short descriptions of structured nodes which consist of slot/value pairs. There are many publications which deal with learning a generation model from parallel data. These parallel data consist of the structured data and natural-language text, so that the model can learn to transform the structured data into text. Konstas and Lapata (2012) compare 1-*best* vs $k$-*best* approaches in the verbalization of database records; the $k$-*best* is implemented as hypergraph decoding under a trigram language model. Duma and Klein (2013) generate short natural-language descriptions, taking structured DBPedia data as input. Their approach learns text templates which are filled with the information from the structured data. Mei et. al. (2015) use recurrent LSTM models to generate text from facts given in a knowledge base.

Several recent papers tackle the problem of generating a one-sentence introduction for a biography given structured biographical slot/value pairs. Lebret et al (2016) introduced a neural model for this concept-to-text generation and evaluated on a large dataset of biographies from Wikipedia. Chisholm et. al. (2017) solve the same problem by applying a machine translation system to a linearized version of the pairs. All of these approaches require a large set of parallel training data to learn from.

In the work presented here, however, we need to generate titles also in languages for which we do not have parallel (slot/value pairs to natural language) training data. For the romance languages (French, Italian and Spanish), no curated titles are available for training. For these languages, we resort to rule-based language generation. These systems are time-consuming and require significant human effort, but a lot of research work has been done in this area such as the FoG system (Goldberg et al., 1994), the SumTime system (Reiter et al., 2005) and the PLANDOC system (McKeown et al., 1994).

For the languages English and German, we have parallel data available, so we can directly learn with a machine translation (MT) approach to “translate” from rule-based generated titles to curated titles. Note that, both source and target language are identical in our case. This problem is widely studied in the MT community under the umbrella of Automatic Post-Editing (APE) (Simard et al., 2007). APE systems are mostly used to correct the output of a traditional MT system (with different source and target language), thereby producing higher quality translations. To the best of our knowledge, there is no work in the prior art that leverages an APE system to improve the quality of a rule-based generation system.

Another difference between our work and some of the papers described above, (Konstas and Lapata, 2012), (Mei et al., 2015) and (Chisholm et al., 2017), is that they perform selective generation, i.e. they run a selection step that determines the slot/value pairs which will be included in the verbalization. For our e-Commerce browse pages, all slot/value pairs are relevant and need to be verbalized.

The work of (Zajic and Dorr, 2002) is related in that they add morphological variation of verbs into a HMM approach for headline generation. In our hybrid approach described in section 4, we also incorporate many alternative lexicalizations and let the decoding process find the optimal sequence. However, our work is different in that we generate titles from slot/value pairs instead of news stories, and we allow for much more variation than those used in (Zajic and Dorr, 2002).

## 3 Rule-based approach

The first title generation system described in this paper is a strictly rule-based approach with a manually created grammar. These approaches are especially useful when the amount of human-curated training data is limited. Since on an e-Commerce site, different categories will have different possible slots and slot-value pairs, the total number of potential slots can be huge. In this case, creating individual rules for each slot is not feasible. But we can heuristically classify slot/value pairs into a small set of slot types. With Table 1 as an example, we classify

- All slot/value pairs with an adjective value as *Adjective slots*, e.g. *Color: white*
- All slots mentioning a brand, model, series, maker as *Brand slots*, e.g. *Brand: ACME*
- All slots with a numerical value as *Numerical slots*, e.g. *Storage Capacity: 32GB*
- All slots with a Boolean (Yes/No) value as *Boolean slots*, etc.

Furthermore, in the slot classification phase, we can use a language model trained on related-domain texts (product titles and description, search queries, . . . ) to identify whether certain nominal aspect values typically go with specific prepositions, e.g. "*in* English" for (*Language: English*) slots, "*for* children" for (*Age group: Children*) slots.

Each of these slot types then gets a hand-written language-dependent rule for lexicalization[1], using parts of the slot value as well as the slot name, if required: For example, adjective slot values are inflected to match the category head noun inflection[2] and realized by themselves (*Category: Schuhe*) + (*Farbe: Grün*) → "grüne (Schuhe)"[3]. Numeric aspects are realized as combination of an optional preposition, the slot name including units, and the numeric value (*Diamètre (cm): 20*) → "diamètre 20 cm"[4]). Certain slots can also replace parts of the category name; e.g. (*Category: RC Boats & Watercrafts*) + (*Type: Submarine'*) is combined into "RC Submarines".

We then create a language-dependent grammar which combines all these realizations by slot type in a defined order, e.g. for English, BRAND | NUMERIC | ADJECTIVES | NOMINAL | CATEGORY+TYPE | FROM | WITH | FOR | LANGUAGE | BOOLEAN.

## 4 Hybrid approach

### 4.1 Motivation

This section describes the hybrid generation approach which combines the rule-based language generation approach (Section 3) and statistical machine translation for situations in which monolingual data for the language is available, but human-curated titles are not.

The approach described in Section 3 requires creating and maintaining individual rules for all potential slots in all categories for all languages, which is next to impossible on a large e-Commerce site. The structure of categories and slots is dynamic and typically evolves over time. Furthermore, there are many combinations of slots that lead to redundancies and non-fluent generations, which are hard to cover with individual rules. For example, if we have a "Room" slot, any mentions of "interior-" in a title become redundant. We therefore developed a model

[1] Lexicalization is the same as verbalization or realization in this context.

[2] We use a shallow tagger for that

[3] "green shoes"

[4] "diameter 20cm"

which is able to learn and generalize the realizations from data.

### 4.2 Statistical machine translation

The process of generating a title from category information and slot/value pairs can be modeled by leveraging phrase-based statistical machine translation (PBSMT). Translation from source into target language using PBSMT works as follows:

1. the source sentence is split into all possible sequence of words called *phrases*[5],
2. each of these source phrases is looked up in a phrase translation table (learned from the training data) and translated into a target phrase,
3. these target phrases are then combined to form a translation candidate,
4. a language model scores the translation candidates, and
5. the system outputs the best scoring translation.

We use the open-source Moses translation system (Koehn et al., 2007) for our hybrid system. In order to combine the rule-based generation approach with machine translation decoding, we leverage the so-called cache-based translation model (Bertoldi et al., 2014). This model uses an additional dynamic phrase table providing one score per phrase pair, originally implemented for dynamic adaptation. CBTM has been integrated into Moses, and we extended the available implementation to match our use case of title generation:

- CBTM is intended to work at the document level, while we are working on the title level. We extended this so that each title has a separate cache model which is not accessible to any other thread but the thread assigned to that sentence, thus also making the model thread-safe;
- The original CBTM score decays over time, i.e. the recent phrase pairs are scored higher than the old phrase pairs, where recency refers to document history. We modified this so that the cache entries do not decay over time. Instead, the cache entries are deleted once the sentence is processed;
- CBTM generates only one score based on recency of the phrase pair. We extended this multiple static scores for each cache entry, similar to a standard phrase-table.

| Slot/*Value* | Lexicalizations |
|---|---|
| Category/*Cell Phones & Smart Phones* | ”Cell Phones & Smart Phones” |
| Brand/*ACME* | ”ACME” |
| Color/*white* | “white”, “in white” |
| Storage Capacity/*32GB* | “32GB”, “with 32GB”, “with storage capacity 32GB”, “, 32GB”, ... |

**Table 2:** Alternative lexicalizations for the slot/value pairs for the browse page in Figure 1

### 4.3 Combining rule-based and PBSMT

The rule-based generation approach generates exactly one lexicalization for each slot/value pair. In the hybrid approach, we extend this and generate many different alternative lexicalizations[6] for a slot/value pair.

For the example browse page from Table 1, alternative lexicalizations for the slot/value pairs are shown in Table 2. Note that we also add the category as a slot and the category name as its lexicalization. This is done in order to generate distinct title for products that have same slot/value pairs across categories.

We use the cache-based translation model described in section 4.2 to represent these alternative realizations. The slot/value pairs are represented as the source phrases, and their alternative lexicalizations are represented as different target phrases, i.e. different possible ”translations” of this source phrase. For each browse page, we dynamically create a specific phrase table containing source representations and their possible target phrases.

These target phrases are then combined and scored by a language model, and the system returns the best scoring title. This language model

[5] These are sequences of contiguous words and not necessarily linguistically well-defined phrases.

[6] We use a shallow tagger to lexicalize the slot names and values.

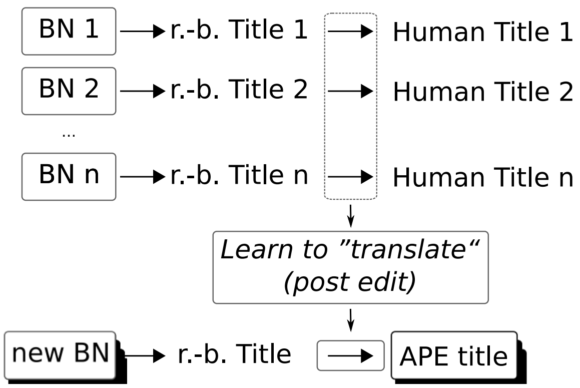

**Figure 2:** Principle of Automatic Post Editing. "r.-b. Title" denotes an auto-generated (rule-based) title as described in Section 3.

is trained on a large corpus of monolingual out-of-domain data. As mentioned in the introduction, we have set up the hybrid system for languages where no curated browse page titles are available. However, we do have access to monolingual data which can be used to build a language model.

## 5 Automatic Post-Editing approach

Rule-based title generation causes some errors in the output which are recurrent and consistent. For English and German, we have access to a reasonably large amount of already available curated titles (refer Table 3). By generating titles using the rule–based system for the same browse pages, we can create parallel data, where the source side is the – sometimes erroneous – automatically created title, and the target is the – desired human–curated title. Note that in this case, both source and target text are in the same language.

These parallel data can be used to train an automatic post-editing (APE) system (Simard et al., 2007) which will then learn how to correct the errors made by the rule-based system. In this work, we are learning to automatically correct the titles generated by the rule-based system. As proposed in APE, we train a phrase-based statistical MT system on the "parallel" data where the source corpus is the rule-based generated data and the target corpus is the human curated title data. The principle behind this approach is sketched in Figure 2. We use the Moses toolkit (Koehn et al., 2007) for training and decoding.

The APE approach has several advantages: it is straightforward to implement, automatic post-editing is a well-studied topic, and we can apply all features and models used within the current state-of-the-art MT systems.

We leverage the following MT models:

1. Phrase translation model, including a simple string-matching penalty that is used to control for higher faithfulness with regard to the raw machine translation output (Junczys-Dowmunt et al., 2016). This penalty is added as phrase score in the phrase translation model.

2. Operation Sequence Model (OSM): This models the translation process as a Markov chain of sequence of operations. OSM basically clubs the benefits of both translation and language model into one model (Durrani et al., 2011).

3. Language Model (LM): The 5-gram LM was trained only on the target side of our "parallel" data, i.e. the human-curated titles, because this is clean in-domain data.

At the same time, there are some disadvantages of using an APE system:

- Only errors seen in training can be corrected by APE;

- When the data is noisy, APE learns to generate artifacts from the data.

- Data that is lost in the rule-based generation cannot be reconstructed.

## 6 Experiments

### 6.1 Data

We evaluated our rule-based, HybridMT and APE approach for the generation of browse page titles in English (En), German (De), French, Italian and Spanish (FrItEs). The rule-based system does not require any training data. The HybridMT system requires a language model (LM) which we trained on large monolingual data extracted from descriptions of items in the e-Commerce inventory. For FrItEs, we do not have access to many human-curated titles (we have access to small development and evaluation sets), but we do have these item description data available. We heuristically filter the description data by making sure that each sentence has at least one preposition and the number of tokens are greater than 3. Parallel data is used to train, tune and evaluate the APE system.

For development and testing purposes, we use a manually-generated set of 500 browse page titles per language. Statistics on the amount of parallel and monolingual data for the languages are given in Table 3. "AlgoTok" represents the number of tokens in browse page titles which were generated using the rule-based system, while "CuratedTok" represents those in human curated titles.

| | Curated | | Monolingual |
|---|---|---|---|
| Language | #SrcTok | #TrgTok | #TrgTok |
| English | 5.21M | 4.97M | 4.97M |
| German | 3.68M | 3.76M | 3.76M |
| French | - | - | 47.3M |
| Italian | - | - | 39.8M |
| Spanish | - | - | 56.7M |

| | Dev | | Test | |
|---|---|---|---|---|
| Language | #Src | #Trg | #Src | #Trg |
| English | 7.5K | 6.7K | 6.7K | 6.6K |
| German | 8.5K | 8.8K | 8.6K | 8.8K |
| French | - | 3.2K | - | 3.8K |
| Italian | - | 10K | - | 3.7K |
| Spanish | - | 6.7K | - | 3.6K |

**Table 3:** Statistics of training (Curated, Monolingual), development (Dev) and evaluation (Test) datasets. For French, Italian, Spanish there is no parallel training data. M, K stands for million and thousand respectively.

### 6.2 Systems

This section describes the various language generation systems we have applied. All the language models are 5-grams with modified Kneser-Ney smoothing trained with KenLM (Heafield, 2011). We use the modified cache-based translation model (cf. Section 4) for the HybridMT systems. For the APE system, we train the translation and operation sequence model with scripts provided under Moses (Koehn et al., 2007). For tuning the weights we use the k-best batch MIRA implementation (Cherry and Foster, 2012) provided in the Moses toolkit. A combination of BLEU (Papineni et al., 2002) and word error rate (WER) (Nießen et al., 2000) is used for tuning the system, because tuning on BLEU only resulted in overly long translations. Performance of all systems are reported in terms of BLEU, character F1-score CHRF1 (Popović, 2016) and WER. Statistical significance tests were conducted using approximate randomization tests (Clark et al., 2011).

## 7 Results

This section collects the results from the three different generation systems, namely the rule-based system (RBNLG), the Hybrid Machine Translation system (HybridMT) and the automatic post-editing system (APE), for 5 different languages. We group the results based on the amount of curated titles available per language. For English and German, we have curated titles and can thus train an APE system. For FrItEs, we do not have such data, so we compare the RBNLG system with the HybridMT approach. In addition, we also run a contrastive experiment on English with the HybridMT system, comparing the quality for all three systems on this language.

### 7.1 English and German

Table 4 collects results for English and German. In general, the results obtained by the APE systems are significantly better than the rule-based system. For English, the APE system shows an absolute improvement of almost 10 BLEU points over the rule-based system. The effect of in-domain data can be clearly seen in this experiment.

Our rule-based approach for German in comparison with English does not fare well. German is a morphologically rich language and to manually cover all the grammar rules in German requires a lot of effort. This is one of the reasons why the metric scores on German are far lower than on English. These errors by the rule-based system, such as reordering errors, usage of prepositions etc., are however systematic and consistent across titles. These consistent errors are captured well by the APE system which results in a very impressive improvement: the APE system outperforms the rule-based approach by almost 30 BLEU points and reduces WER by 29% absolute.

The HybridMT system on English is also better than the RBNLG system by almost 3 BLEU points. In this particular case, HybridMT benefits from the large monolingual in-domain language model trained on the curated titles. The tuned weight of the in-domain LM is 0.41 in comparison to the out-of-domain LM with a weight of 0.13.

| Language | System | BLEU | CHRF1 | WER |
|---|---|---|---|---|
| English | RBNLG | 69.96 | 87.30 | 25.82 |
| | HybridMT | 72.71 | 88.44 | 22.25 |
| | APE | 80.29 | 91.71 | 15.89 |
| German | RBNLG | 41.68 | 76.68 | 56.00 |
| | APE | 65.10 | 89.6 | 29.51 |

**Table 4:** Cased BLEU, Character F-score and WER on the human post-edited data for English and German.

### 7.2 FRITES

Table 5 collects results for French, Italian and Spanish. We see a similar trend for these three languages as we saw for English and German. In all cases, the HybridMT system is significantly better than the RBNLG system. The largest gains are observed in French and Italian (7–10 BLEU points) and relatively moderate improvements in Spanish (∼4.5 BLEU points).

| Language | System | BLEU | CHRF1 | WER |
|---|---|---|---|---|
| French | RBNLG | 66.47 | 86.79 | 27.20 |
| | HybridMT | 73.32 | 89.42 | 22.87 |
| Italian | RBNLG | 49.92 | 79.08 | 38.28 |
| | HybridMT | 60.63 | 83.67 | 30.98 |
| Spanish | RBNLG | 65.33 | 85.83 | 26.57 |
| | HybridMT | 69.92 | 87.40 | 23.14 |

**Table 5:** Cased BLEU, Character F-score and WER on the human post-edited data for French, Italian and Spanish.

### 7.3 Analysis of HybridMT

We have seen that HybridMT takes advantage of the alternative phrase pairs generated in the output of the rule-based system and then leverages the LM to score the title. With this system, we can also generate an n-best list of titles for a particular browse page. The HybridMT system picks the best title as scored by the decoder. This one title might indeed be the best title the system can generate. However, there might be even better alternatives which the system can generate, but which receive a worse score from the decoder models.

To find this out, we did several experiments with the HybridMT system for the FrItEs languages. We looked for the best possible title in the n-best list by comparing sentence-level BLEU scores of the candidates in the n-best against the curated title. This gives us the upper bound of the quality of translation that can be achieved. With the increasing size of the n-best list we found out that we can find better generated titles. Figure 3 plots the BLEU scores against the increasing size of n-best list respectively. As we can observe in these figures, BLEU score continuously increases with the increasing size of n-best and plateaus after a while.

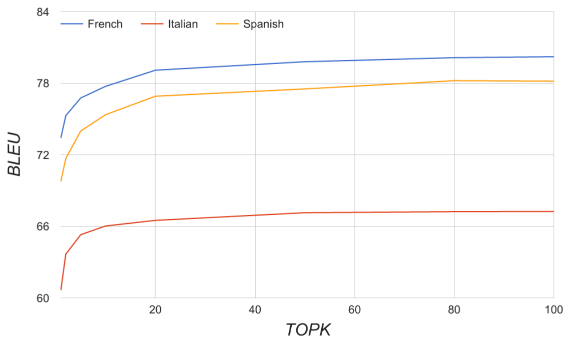

**Figure 3:** Learning curve of HybridMT system with increasing size of n-best list against BLEU score.

This experiment basically shows that there is a large scope of improvement in the quality of titles if we have access to a large amount of in-domain data or curated titles in these languages.

### 7.4 Qualitative Evaluation

In Table 6 we show examples from the evaluation set where our HybridMT and APE systems solved the problems occurring in the rule-based RBNLG system.

One of the most common issues we encountered in the RBNLG system is redundancy. In this rule-based system, all slot/value pairs are realized independently of each other in the title. This realization approach can cause redundancy in the title if there is overlap between two slot/value pairs. For example, consider the browse page with slot/value pairs (*Category: Lamps*, *Type: Side lamps*, *Color: Blue*, *Style: Decorative*). The RBNLG system will generate the title “Decorative blue lamps side lamps”, whereas the reference title is “Decorative blue side lamps”. This happens because RBNLG system does not know that side lamps are also lamps.

| Language | System | Title |
|---|---|---|
| **English** | RBNLG | Album Suite Classical SACD Music CDs |
| | HybridMT | SACD Album Suite Classical Music CDs |
| | APE | Album Suite Classical Music SACDs |
| | Ref | Album Suite Classical Music SACDs |
| **German** | RBNLG | Ab 12 LEGO Baukästen & Sets mit Weltraum |
| | APE | LEGO Weltraum-Baukästen & - Sets ab 12 Jahren |
| | Ref | LEGO Weltraum-Baukästen & - Sets ab 12 Jahren |
| **French** | RBNLG | Réflecteurs pour automobiles Marque du véhicule BMW |
| | HybridMT | Réflecteurs pour automobiles BMW |
| | Ref | Réflecteurs pour automobiles BMW |

**Table 6:** Examples of English, German and French titles generated by rule-based, HybridMT and APE systems.

A similar problem occurs in the first English example in Table 6: RBNLG realizes all slot/value pairs independently, and ends up generating both "SACD" (Super Audio CD) and "CDs" in the title. The same happens in the HybridMT system, which generates a title containing both "SACD" and "Music CDs". The automatic post-editing system, on the other hand, learns to fix this redundancy in the title and drops "CDs" and at the same time inflects "SACD" to its plural form.

While realizing the slot/value pairs, the rule-based system has an option to either verbalize the "slot" or drop it and just output the "value". This is done in order to not generate any redundant information in the title. In the German example[7] in Table 6, we see that the RBNLG system incorrectly dropped the slot and did not verbalize it. For the slot/value pair (*Jahre: Ab 12*), RBNLG generates the incomplete verbalization "Ab 12" and puts it in the wrong position within the title. APE learns to correct these kind of errors from the parallel data, and fixes both the ordering and the missing word.

The French titles[8] generated by RBNLG and HybridMT in Table 6 are another example of the same issue. In this case, the slot/value (*Marque du véhicule: BMW*) is realized as "Marque du véhicule BMW" by the RBNLG system, i.e. both slot and value are output. This is not needed since "pour automobiles" already contains the information about the "véhicule". In HybridMT, the cache-based translation model has the option of realizing the slot/value pair as "Marque du véhicule BMW" or "BMW", and in this case the language model chooses the latter option, thereby improving over the rule-based system.

## 8 Conclusion

We have described three different approaches for automatic generation of browse page titles. The rule-based approach can be applied on languages for which we do not have in-domain data at all. The hybrid machine translation approach extends this approach and uses monolingual in-domain data, yielding substantial gains over the rule-based system. For settings where we have large amounts of human-curated browse page titles already, we developed an automatic post-editing system which can be applied on top of a rule-based system and leads to very impressive improvements.

In future work we are planning to extend the hybrid approach to German by adding the generation of realization alternatives, and compare this with the APE approach. We will also investigate on combining learned and rule-generated realizations, and on using different Machine Translation approaches for APE and the Hybrid model. We will also work on generating the browse page titles directly from the realized form of meta-data (i.e. concatenation of all slots and slot-values pairs) using a sequence to sequence model with attention mechanism as described in Bahdanau et. al. (Bahdanau et al., 2014).

## Acknowledgments

We would like to thank our colleagues Alex Zhicharevich and Daniel Hurwitz for developing the initial rule-based approach for English and German that was the foundation of this work.

[7] Translation: LEGO Space kits & sets from 12 years

[8] Translation: Reflectors for BMW cars

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Nicola Bertoldi, Patrick Simianer, Mauro Cettolo, Katharina Wäschle, Marcello Federico, and Stefan Riezler. 2014. Online adaptation to post-edits for phrase-based statistical machine translation. *Machine Translation*, 28:309–339.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June. Association for Computational Linguistics.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 633–642. Association for Computational Linguistics.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Robert Dale, Stephen J Green, Maria Milosavljevic, Cécile Paris, Cornelia Verspoor, and Sandra Williams. 1998. The realities of generating natural language from databases. In *Proceedings of the 11th Australian Joint Conference on Artificial Intelligence*, pages 13–17.

Daniel Duma and Ewan Klein. 2013. Generating natural language from linked data: Unsupervised template extraction. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 83–94. Association for Computational Linguistics.

Nadir Durrani, Helmut Schmid, and Alexander M. Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1045–1054, June.

Eli Goldberg, Norbert Driedger, and Richard I. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and Their Applications*, 9(2):45–53, April.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The AMU-UEDIN submission to the WMT16 news translation task: Attention-based NMT models as feature functions in phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325, Berlin, Germany, August. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Ioannis Konstas and Mirella Lapata. 2012. Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 369–378. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Generating text from structured data with application to the biography domain. *Computing Research Repository (CoRR)*, abs/1603.07771.

Kathleen McKeown, Karen Kukich, and James Shaw. 1994. Practical issues in automatic documentation generation. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, ANLC '94, pages 7–14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2015. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. *Computing Research Repository (CoRR)*, abs/1509.00838.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Language Resources and Evaluation (LREC)*, pages 39–45, Athens, Greece, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation*, pages 499–504, Berlin, Germany, August. Association for Computational Linguistics.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, NAACL*. The Association for Computational Linguistics, April.

David Zajic and Bonnie Dorr. 2002. Automatic headline generation for newspaper stories. In *Proceedings of the DUC 2002 workshop on text summarization*, July.