Natural Transmission of Information Extraction Results to End-Users – A Proof-of-Concept Using Data-to-Text*

José Casimiro Pereira[1], António J. S. Teixeira[2], Mário Rodrigues[3], Pedro Miguel[4], and Joaquim Sousa Pinto[5]

1 **Politecnic Institute of Tomar (IPT), Tomar, Portugal**
casimiro@ipt.pt
2 **Department of Electronics, Telecommunications and Informatics/IEETA, University of Aveiro, Aveiro, Portugal**
ajst@ua.pt
3 **ESTGA/IEETA, University of Aveiro, Aveiro, Portugal**
mjfr@ua.pt
4 **Department of Electronics, Telecommunications and Informatics/IEETA, University of Aveiro, Aveiro, Portugal**
5 **Department of Electronics, Telecommunications and Informatics/IEETA, University of Aveiro, Aveiro, Portugal**

Abstract

Information Extraction from natural texts has a great potential in areas such as Tourism and can be of great assistance in transforming customers' comments in valuable information for Tourism operators, governments and customers. After extraction, information needs to be efficiently transmitted to end-users in a natural way. Systems should not, in general, send extracted information directly to end-users, such as hotel managers, as it can be difficult to read.

Naturally, humans transmit and encode information using natural languages, such as Portuguese. The problem arising from the need of efficient and natural transmission of the information to end-user is how to encode it. The use of natural language generation (NLG) is a possible solution, for producing sentences, and, with them, texts.

In this paper we address this, with a data-to-text system, a derivation of formal NLG systems that use data as input. The proposed system uses an aligned corpus, which was defined, collected and processed, in about approximately 3 weeks of work. To build the language model were used three different in-domain and out-of-domain corpora. The effects of this approach were evaluated, and results are presented.

Automatic metrics, BLEU and Meteor, were used to evaluate the different systems, comparing their values with similar systems. Results show that expanding the corpus has a major positive effect in BLEU and Meteor scores and use of additional corpora (in-domain and out-of-domain) in training language model does not result in significantly different performance.

The scores obtained, combined with their comparison with other systems performance and informal evaluation by humans of the sentences produced, give additional support for the capabilities of the translation based approach for fast development of data-to-text for new domains.

1998 ACM Subject Classification I.2.7 [Natural Language Processing] Language Generation, Machine Translation, H.5.2 [User Interfaces] Natural Language

Keywords and phrases Data-to-Text, Natural Language Generation, Automatic Translation, opinions, Tourism, Portuguese

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.20

* Research partially funded by IEETA Research Unit funding (UID/CEC/00127/2013) and Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

6th Symposium on Languages, Applications and Technologies (SLATE 2017).
Editors: R. Queirós, M. Pinto, A. Simões, J. P. Leal, and M. J. Varanda; Article No. 20; pp. 20:1–20:14

Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

To: *Director of HOTEL XYZ Aveiro*
Subject: March 2017 evaluations

Your hotel had a mixed review in restaurant service AND bad at room service. The price is high. Your competitor, Hotel YYY, had good evaluation at restaurant service AND very good at room service. Their price was considered good.
2 April 2017 *[Automatically generated by IE + NLG]*

Para: Director do Hotel XYZ de Aveiro
Assunto: Avaliacões do mês de março 2017

O seu Hotel teve avaliação mista nos serviços de restauração E péssima nos quartos. O preço é alto. O seu concorrente, Hotel YYY, teve avaliação boa nos serviços de restauração E muito boas nos quartos. O preço dele foi classificado como bom.
2 de Abril 2017
[Gerado automaticamente por EI + GLN]

Figure 1 Example of hypothetical email with hotels' evaluations. English version at the top.

1 Introduction

User comments about service experiences have a key role in influencing the consumption decisions of other customers. The widespread adoption of Internet technologies expanded enormously the volume of such comments and the potential impact of such customer generated content. An area where user opinions are particularly influential is Tourism, as hotels, restaurants, and attractions are constantly commented by users.

For service providers, such as hotel owners and managers, this feedback provided by customers is very important, but it is impossible to even grasp the information in all this comments by manual processes. Information Extraction (IE) from natural texts (e.g. [12]) can be of great assistance in processing customers comments and extracting relevant information, but systems cannot, in general, send extracted information unfiltered to an end-user, such as an hotel manager, as the raw information can be hard to understand. Gathered information needs to be efficiently transmitted to end-users in a natural way.

In this paper we address this problem, of efficient and natural transmission, with a data-to-text system, a derivation of formal Natural Language Generation (NLG) systems that use data as input (e.g. log events). The application scenario, selected for the development of a proof-of-concept, consists of having the data-to-text system as a "translator" to Portuguese sentences of the information extracted by IE from comments to hotels' services, made in Portuguese by their customers. IE provides information such as names of hotels, services, classifications given to services and the amount of people who classified services. The target, when system becomes completed, is the production of short text, expressing the opinions that customers give about hotels' services. This text, for example, could be sent by email to hotel managers each morning. Figure 1 presents a vision of such email.

This paper is organized as follows: in the next section, related work is presented, followed by an overview of the structure of proposed Data-to-Text system. Section 4 presents the scenario of application, and how the system was adapted to the domain. In Section 5 several examples are presented of system output and the evaluation of system variations. A comparison with similar systems was made too. The paper ends with conclusions, and acknowledgements.

2 Related Work (in Data-to-Text)

Reiter [10] defined *Data-to-Text* systems as *systems that generate texts from non-linguistic input data, which is typically numerical.* The major difference between 'classic' NLG systems

and Data-to-Text systems is that this last category must analyse and interpret their input data, whereas the other category does not need. Of course, they need, as well, to decide how to linguistically communicate the produced utterance. This section presents, briefly, some representative Data-to-Text systems, by chronological order, with focus on the ones employing data driven approaches.

Pollen Forecast for Scotland system. Pollen Forecast for Scotland system [15] was one of first Data-to-Text systems to be reported. Its aim is to report, on text, the prediction of pollen concentration on different regions of Scotland. It was made of two related sub-tasks: the prediction itself, and the translation of numerical data to text, provided by prediction. The translation task is based on a parallel corpus of 69 data-text pairs. Each pair corresponds to a pollen concentration data and the corresponding forecast. All forecasts were written by expert meteorologists.

To be able to generate correct texts, from meteorological data, all human written forecasts were analysed in respect to three dimensions: Message type; Data dependency and Corpus coverage. In addition, input data was analysed in three steps: segmentation of geographic regions by their non-spatial attributes (pollen values); segmentation of each segmented geographic regions by their spatial attributes (geographic proximity); and, detection of trends in the generalized pollen level for the whole region over time. With these segmented data, Pollen Forecast system produces vectors of trends for pollen prediction. These vectors were used to determine the correct forecast text to be produced.

BabyTalk. BabyTalk project [9] works with data from a Neonatal Intensive Care Unit (NICU), from a Scottish hospital. It tends to join two concepts: raw medical data and recommendation of specific actions to the medical staff. The aim is to produce a text summary, produced in Natural Language, where data and instructions should be put together.

BabyTalk has five different sub-projects, each one related with a particular issue of NICU centre, including: *BT-Nurse*, designed to automatically generate English summaries of the electronically recorded patient data over a twelve hour nursing shift; and *BT-45*, intended to present reports to nurses and doctors that summarize 45 minutes of patients' data.

BT-45 was the first to be implemented and is a truly data-to-text system. However, instead of only having numerical data as input, as weather reporting systems, BT-45's input is more heterogeneous. Most data comes from a NICU database, and includes an ontology that is used both to represent domain knowledge and to support linguistic processing. The BT-45 follows the data-to-text architecture suggested by [10], hence the presence of a Document Planning, Micro-Realization and Realization modules.

Mountain. Mountain was developed by Langner for his PhD [4]. Mountain was the only of the analysed systems that uses the Moses[1] tool. This statistical tool is used to train and process a parallel aligned corpora. The translation model built by Moses is used by Mountain to generate an output sentence, translated from input language.

Mountain's scenario was reservations of tennis court, where users make reserves from one to several hours. An aligned corpus was produced, with two languages. One with answers given to reservations, by English native speakers, and another with a set of three tokens expressing the requested scheduling. On the second language, the first token expresses the

[1] http://www.statmt.org/moses/

Table 1 Example of MOUNTAIN's output. Left column: *input* sentence; right column: generated sentences, from [4, pag.71].

000000 d5 t3	friday evening is completely closed
100000 d2 t2	the only time available is noon
111111 d4 t1	the court is open all morning
111111 d1 t3	you can reserve a court anytime on monday evening
100011 d5 t3	six , ten or eleven

court availability. Each number represents one hour, over a day. `1` is for 'vacancy', and `0` is for 'taken'. Second token expresses the day of week, and last token represents the day period – morning, afternoon or evening. `111111 d2 t3` and `001110 d5 t1` are examples of it.

The collected corpus was expanded, from 800 to almost 4500 sentences, because it was realized that most sentences, produced by humans, with some minimal changes, could be reused, without losing their original meaning. After that, the corpus was trained with MOSES tools. The training model produced was responsible for translations carried out by MOUNTAIN.

Table 1 presents a sample of produced output sentences by MOUNTAIN. Some produced sentences were directly translated from *output corpus*, likely due to the presence of direct input–output pair on corpus. Others were the combination of two or more original sentences. Overall, more than three quarters of the generated responses are not present in the original corpus.

SINotas. SINotas [5] is one of few, for the best of our knowledge, data-to-text system that uses Portuguese language. In this case Brazilian Portuguese. The goal of SINotas is the production of short textual reports of students grades, weekly attendance rates, and other related academic information.

SINotas has a small corpus of 241 paired data-text records of students' academic performance [5]. All pairs were related to students of a single professor. The same professor was responsible for writing the text that expresses the contents in the the data. The use of a single professor, defined as 'domain expert', was justified with the need of coherence on meanings from raw data (e.g., students' grades) to semantics (i.e., the interpretation of the data according to a professor). The use of different 'domain experts' could generate contradictory description for the same data.

All output texts, that describe raw data, are up to five sentences long, offering students a description about their grades in relation with other students' grades.

PortNLG. PortNLG [14] is a recent tool that processes the last phase of NLG pipeline defined by Reiter [11] – the Surface Realization, on Portuguese language. PortNLG is a Java library, developed to be integrated on data-to-text systems, hence is not itself a real NLG system. It works similarly to SimpleNLG [3]. Like SimpleNLG, it uses grammar rules, but, since Portuguese grammar is more complex than English grammar, PortNLG needs an extra component: a lexicon module. The application where PortNLG is integrated, executes a sequence of calls to PortNLG methods, constructing, this way, the desired sentence.

Medication2PT. MEDICATION2PT [7] is a recent Data-to-Text work made to provide instructions in European Portuguese language about taking medicines. This system was

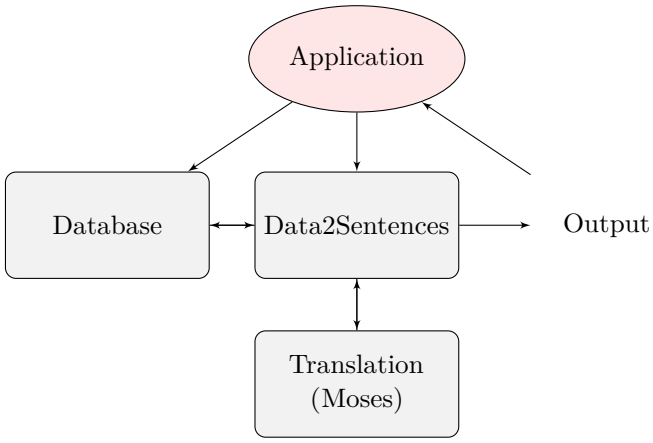

Figure 2 Architecture of the Data-to-Text developed system.

integrated in a larger system, the Medication Assistant [2] designed to take advantage of Smartphones, and targeting elderly users.

The MEDICATION2PT was inspired by MOUNTAIN's work. It uses MOSES as main engine to process the input data and obtain output sentences. Since its intention is to build a system with low resources, it uses a small corpus to train MOSES. As a consequence, a small part of produced sentences has unacceptable quality, and can not be used. To identify sentences with lower quality, authors proposed a classifier module capable of providing information on the Intelligibility or Quality of the sentences. Sentences marked as unacceptable are replaced by template-based generated ones. This classifier module combines extraction of linguistic features with a classifier trained in a manually annotated corpus [8].

Later, in 2015, MEDICATION2PT classifiers and templates modules were integrated in a Hybrid System, proposing an integrated solution to produce sentences in Portuguese language, at a lower cost [8].

3 The Developed System

The developed system is based on the use of machine translation and draws on recent works, such as the MOUNTAIN and MEDICATION2PT, from the authors. The overall architecture of the developed system is presented in Figure 2.

The central part of this system is a module, named **Data2Sentences**, able to create sentences in response to vectors with data, provided as input. If several input vectors are sequentially fed to the Data2Sentence module, an aggregation module can be used to join the set of generated sentences to produce a small text.

The **Database** (DB) **module** is the component responsible for storing all data related with system where this module is used.

The **Translation module**, based in MOSES translation system, is responsible for the translation to Portuguese of the input vectors sent by the Data2Sentences module. To perform this task, MOSES must be trained with a corpus consisting of two languages perfectly aligned. Every sentence from first language, the *input language*, must match a phrase in the second language, the *output language*. In our system the input language consists of values related with the ones present on the DB module. The output language is constituted by expressions in Portuguese that represent the data on input language. MOSES uses this relation to generate new sentences, when an input vector is provided. Besides the definition

Table 2 *Tokens* of input language and their correspondences.

Token	Correspondence
Hotel name	name of hotel, under evaluation
Service name	name of hotel's service, that is going to be evaluated
Evaluation	evaluation to service
Number people	amount of people who did the evaluation
First adjective	main adjective that reflects evaluation
Second adjective	secondary adjective

of languages and corpora, training Moses implies the creation of a language model. Because words in sentences do not combine in random order, the language model establishes the probability of one or more words appearing in a sentence, in a particular order. These probabilities are used in the process of translation. Commonly, language models consist of `n-grams` with probabilities, estimated from a training corpus of sentences on the language that Moses is expected to translate.

The **Data2Sentences module** is responsible for the coordination of all process. It receives requests from users and interacts with the data stored in the Database. It is also responsible for sending messages (written in the input language) to the Moses-based Translation module and receive its answer, in the output language. Finally, it is in charge of processing the answers and transmit them to the Application.

4 Corpora and training

4.1 Corpus structure

The first task is the definition of the input language and corresponding output language. The proposed work is integrated on an ongoing project, related with data extraction. The aim of that project is to collect and summarize data related with classifications to hotels' services, made by their customers. The provided data consists, mainly, in hotels names, service names, classifications given to services and the amount of people who classified the services. The input language is based on data stored on database module. A sentence on this language is a collection of tokens, in a specific order. Table 2 presents the tokens used in input language, and their order. The output language consists of Portuguese sentences, expressing, by words, the data of input language.

The first token corresponds to the hotel's name. The second token is related to the hotel's service name, that is subject to evaluation. The evaluation of the service is expressed on the Evaluation token, in three possible values: positive, negative, or neutral. The fourth token defines the amount of people that made the evaluation. It is configured with five possible values: very few people, some people, half of evaluators, many people, almost all people. Fifth and sixth tokens are related with adjectives used to express the evaluation to hotel's services. In all tokens, data is compressed and white spaces are replaced by hyphens. That is done to ease Moses work, because that way Moses treat the several words of tokens as a single word. To increase the accuracy of Moses, we need to use all tokens, in every input sentence, whether having, or not, data. When there is no data, the token related with that kind of data is identified with the word 'sem-valor' (no-data). A sample of our corpus is presented at Table 3.

Table 3 Sample from aligned corpus ID1. For each example, first the real content (in Portuguese) is presented. After, in italic, English translations are presented.

Input/Internal language	Corresponding Sentence
hotel-sem-valor servico-atendimento aval-neutra algumas-pessoas adj-diferente adj-aceitavel	algumas pessoas classificaram de forma neutra o atendimento, como diferente dos outros hotéis e aceitável
hotel-no-data service-customer-service aval-neutral some-people adj-different adj-acceptable	*some people classified as neutral the customer service, different from other hotels and acceptable*
hotel-faro-vintage-guest-house servico-restaurante aval-positiva poucas-pessoas adj-fantastico adj-sem-valor	Algumas pessoas acharam o restaurante do Faro-Vintage-Guest-House fantástico
hotel-faro-vintage-guest-house service-restaurant eval-positive few-people adj-fantastic adj-no-data	*Some people found the restaurant of Faro-Vintage-Guest-House fantastic*

Table 4 Corpus used in system's evaluation.

Corpus	Num. Sentences	Description
Corpora used to train the system		
ID1 (original)	135	sentences created by humans, in-domain
ID2 (extended)	435	An extended version of ID1 with input language tokens
ID3 (expanded)	2,781	ID2 expanded with new sentences obtained from seeds
Corpora used to enrich the training of the Language Model		
IE	129,130	in-domain – sentences from customers evaluations
OD1	+200,000	out-of-domain – minutes from European Parliament
OD2	+860,000	out-of-domain – CETEMPúblico

4.2 Corpus preparation

After the definition of tokens to be used on input language, is necessary to build the output language. To do that, about 20 people, Portuguese native speakers, with ages from 20 to 60 years old, produced sentences summarising the data presented by tokens. This task was made in two steps. First, we collected 135 sentences. Afterward, a second set of 230 sentences was collected.

With this sentences, several corpora were created to evaluate the system. They are presented in Table 4. First, we named corpus `ID1`, the one with seed sentences produced by humans (technically, we call to these sentences, *in-domain sentences*). This corpus is the *original corpus*, produced by humans. Second, corpus `ID2`, the *extended corpus*, was created, resulting from adding to ID1 all tokens' data used to describe the input language. Third, a corpus `ID3` was made by expansion of ID2. We used the same technique, as described by Langner [4, pag. 69]. This corpus is the *expanded corpus*.

To be used on the creation of *language models*, we defined three other corpora. As corpus `IE` (from, Information Extraction), we joined all the sentences produced by hotels' customers with their reviews. It has about 129 thousand sentences. These sentences belong to the same domain of the sentences produced by the system in evaluation. To test the impact of using *out-of-domain* (OD) sentences, on the language model, the corpus `OD1` was made with

Portuguese minutes of European Parliament[2]. This minutes are part of the Moses for Mere Mortals (MMM) system, a prototype which aims to help the production of a translation chain for real world documents. These minutes have 200 thousand sentences. The last corpus, `OD2`, corresponds to Portuguese sentences obtained from Público newspaper. These sentences belong to Linguateca project – CETEMPúblico [13]. `OD2` corpus has about 860 thousand sentences. Both corpora are, clearly, out of the domain of our system. The goal is to improve the richness of the produced sentences.

The process of collecting corpora by humans took approximately 2 weeks. After that, the corresponding expansion and preparation of corpora for building the language models took another week, leading to a 3 weeks work for all process.

5 Results

5.1 Examples of generated sentences

To give an idea of the generation capabilities of the developed system, Table 5 presents several representative examples of sentences produced. The table is divided in 4 parts.

The first part shows sentences produced by the system for two different input vectors with information from the ID3 corpus. The input vector is presented first, preceded by *in:*; *out:* identifies the generated sentence.

All sentences were produced with the system trained with ID3+IE+OD1+OD2 corpora (see Table 4). It is possible to see that the sentences have distinct quality. The first is considered 'good', since it is grammatically correct and transmits correctly all data expressed in the input vector. The second sentence can be considered 'acceptable' since it transmits the main idea of what is expressed by the input, but it is not grammatically correct nor uses all available data.

The second part of table (sentences 3 and 4) presents an alignment between sentences produced by humans and sentences produced by the system, for the same input. The sentences presented are short to be possible to show differences between human and system sentences. For each input data, when a sentence is produced by system, the generated sentence is usually different from those a human usually do. These differences can be the addition of new words, the suppression of words, or, even, by positioning words in different zones on the sentence. Here, we use the token "***" to highlight the addition or deletion of words. The use of upper cased words is used to represent words that are in different positions on the sentence. Their use on human produced sentences is only to make more visible the differences between the two sentences.

Examples 5 and 6, in part 3 of the table, present good quality sentences produced by the system, both in terms of intelligibility and naturalness. They are completely different from those that humans produced. Nevertheless, they express the same information as the corresponding human sentences. Last part of the table, part 4, presents two sets of sentences generated with different systems. Each set has the same *input* to make possible the comparison of generated sentences. For this set, systems were trained with corpus ID3, ID3+IE and ID3+IE+OD1+OD2. For the first set (example 7), all output sentences can be considered acceptable. They are very different, expressing the same information in quite different ways. The last sentence is the only one grammatically correct. In the last set, example 8, all systems generated the same output sentence.

[2] Available at `https://github.com/jladcr/Moses-for-Mere-Mortals`.

Table 5 Examples of sentences produced by the developed system.

Part 1 – Examples of input vectors and corresponding generated sentences

1	*in*: hotel-no-data service-dinner eval-positive all-people adj-good adj-no-data *out*:o jantar foi bom para a esmagadora maioria das pessoas *(the dinner was good for the majority of people)*
2	*in*: hotel-no-data service-customer-service eval-negative few-people adj-very-bad adj-no-data *out*:atendimento péssimo serviço *(customer service very bad service)*

Part 2 – Examples of produced sentences (S) aligned with the corresponding sentence made by humans (H).

3	*in*:hotel-no-data service-customer-service eval-positive no-data-people adj-pleasant adj-impeccable H: NESTE HOTEL, o atendimento era agradável e BASTANTE impecável S: *** *** o atendimento era agradável e *** impecável
4	*in*:hotel-seaside service-smell eval-neutral no-data-people adj-normal adj-no-data H: NO SEASIDE, o cheiro é DESCRITO COMO normal S: *** *** o cheiro DO SEASIDE é normal

Part 3 – Samples of generated sentences, very different from test sentence, but with quality. H for sentences produced by humans.

5	*in*:hotel-no-data service-noise eval-neutral all-people adj-acceptable adj-no-data H: OUVIA-SE ALGUM BARULHO DE FORA MAS NÃO ERA MUITO INCOMODATIVO S: A ESMAGADORA MAIORIA DAS PESSOAS DESCREVEM O BARULHO COMO ACEITÁVEL
6	*in*:hotel-no-data service-furniture eval-positive few-people adj-pretty adj-functional H: POR ESTAR BONITO E SER FUNCIONAL, DIVERSOS CLIENTES AVALIARAM O MOBILIÁRIO POSITIVAMENTE S: *** *** *** *** DIVERSOS HÓSPEDES GOSTARAM DO MOBILIÁRIO BONITO E FUNCIONAL

Part 4 – Sentences produced by different system variants for the same input.

7	*in*: hotel-no-data service-reception eval-positive half-people adj-charming adj-no-data ID3: a receção cerca de metade dos clientes, como charmosa ID3+IE: a receção , cerca de metade dos clientes charmosa ID3+IE+OD1+OD2: a receção é charmosa
8	*in*: hotel-no-data service-parking eval-positive no-data-people adj-good adj-excellent ID3: o estacionamento é bom e excelente ID3+IE: o estacionamento é bom e excelente ID3+IE+OD1+OD2: o estacionamento é bom e excelente

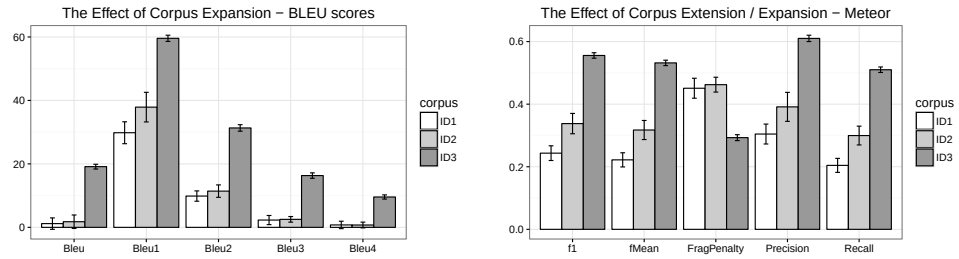

Figure 3 The effect of corpus extension/expansion. BLEU (left) and Meteor (right) evaluation results with the 3 in-domain corpus (original, extended and expanded).

5.2 Evaluation Results

This section presents a formal evaluation made with automatic tools. Several versions of our system were created, using the corpora summarized in Table 4. Evaluation objective was to investigate the influence of the corpus expansion and different language models in the sentences produced, in order to arrive to the best combination.

The commonly used *10-fold cross-validation* technique was adopted, resulting in training 10 different versions (named from A to J) for each variant of the system. After training, all systems were tested with the corresponding test corpus. The BLEU [6] and Meteor [1] metrics were adopted to measure the accuracy of each system.

5.2.1 Effect of corpus expansion

It is well known that Moses works well with corpus of thousands/millions of sentences, but our initial corpus has only 135 sentences. The first questions that arise are: What performance can we have with such a small data set? Can results be improved by methods for extending/expanding? To find answers to these questions, the first evaluation addressed the effect of corpus expansion. Figure 3 presents BLEU and Meteor evaluation results for the 3 in-domain corpus: ID1 (original), ID2 (extended) and ID3 (expanded).

BLEU evaluates the strict correspondence between the evaluated corpus and the reference corpus. With this in mind, the result obtained by ID1 corpus, at BLEU global score is not a surprise. Also, as expected, the improvement to original corpus done with tokens' data (ID2 corpus) provided, globally, better results. The higher improvement was obtained at Bleu1. Effectively, at level of uni-gram the correctness of the provided sentences by ID2 corpus are better than sentences provided by ID1 corpus. This improvement starts to decrease at Bleu2, to a non statistically significant level, once the confidence intervals intersect each other.

The expansion of ID2 corpus with new sentences (ID3 corpus) allows better performance at all levels of BLEU evaluation. The improvement was quite high. BLEU scores corpus from 1 (total correspondence between evaluated corpus and reference corpus) to 0 (no correspondence). The scores obtained with ID3 corpus do not mean that sentences are, globally, not good. It only means that generated sentences are misaligned with reference corpus. And, to be misaligned is not synonymous of bad. As presented at middle section of Table 5, there are good sentences despite their difference from reference sentences.

For Meteor, it is evident that the best evaluation results from ID3 corpus, in all metrics. The F1 metric, which measures, at uni-gram level, the relation between *precision* and *recall*, for ID3 corpus has almost twice the value obtained by ID2 corpus. Almost the same result is obtained with F_{mean} (which favours Recall) with this two corpora. The differences, with

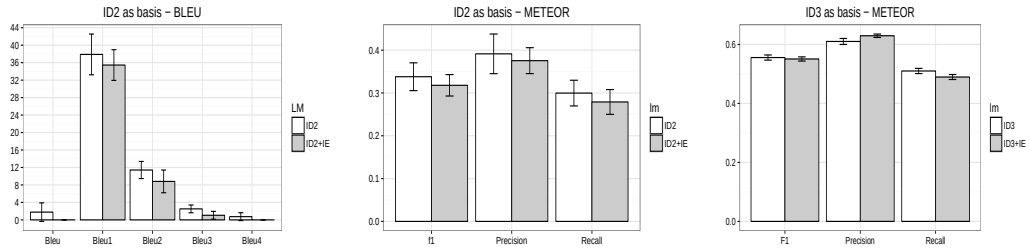

Figure 4 Influence of different Language Models on BLEU and Meteor scores. The first two plots, at left and centre, refer to systems trained with ID2 corpus; plot at right refers for systems using ID3 corpus.

these two metrics, are also statistically significant with corpus ID2 and ID1, although not as large as the difference between ID3 and ID2 corpus.

Equal relation is observed at Precision and Recall metrics, for all corpora. The results for Fragmentation Penalty do not follow the previous tendency. As expected, ID3 corpus has a lower penalty because the corpus is richer. For ID1 and ID2 corpora, the difference is not statistically significant, although ID2 corpus has a slightly higher penalty. That is, probably, due the addition of tokens' data to ID2 corpus, which consist only in uni and bi-grams.

5.2.2 Effect of Language Model training

The second set of questions addressed in the evaluation was related to the effect of the language model, namely: How the corpus used to create the language model for the output language (Portuguese) affects the system performance?

The effect of language model was investigated for systems trained both with ID2 and ID3 corpus. For each, two variants of the language model were trained: one with only the ID corpus; other using ID and IE corpora. Figure 4 summarizes the results obtained. For a first model, only the output language part of ID2 corpus was used in the training process. For the second model, additional corpus IE was used.

Bleu and Meteor average scores obtained with language models trained with only ID2 or ID3 corpora are slightly better than when adding the IE corpus to the training set, but, in general, differences are not statistically significant.

The non-significant difference in scores, confirmed by manual inspection of generated sentences (see examples in Table 6), show as viable the use of both methods to create a language model.

5.2.3 Effect of out-of-domain corpora used in Language Model training

In the previous section, we presented the influence that a richer corpus could have on the production of language models. That enrichment was produced by the addition of a corpus with sentences in the same domain of the output corpus. As out-of-domain corpus are available, the effect of such additional training data for the creation of language models was also evaluated.

Two different corpora were used: (1) a set of Portuguese minutes of European Parliament – named as OD1 corpus; (2) a large set of Portuguese sentences from *Público* newspaper gathered by Linguateca project – named as OD2 corpus. As training corpus, we used ID3 corpus, as previous experiences revealed this corpus can achieve better results.

Table 6 Output samples of systems based on ID2 corpus and using 2 different Language Models: one trained using only ID2 corpus; other trained with ID2 plus IE corpus.

input 1	hotel-no-data service-prize eval-positive few-people adj-accessible adj-no-data
ID2	os clientes do hotel a qualidade de serviço acessível
	in English: The hotel clients service quality accessible
ID2 + IE	a qualidade do hotel é acessível
	in English: The hotel quality is accessible
input 2	hotel-no-data service-wc eval-negative many-people adj-uncomfortable adj-no-data
ID2	o wc a hotel é desconfortável
	in English: the wc the hotel is uncomfortable
ID2 + IE	o wc é muito desconfortável para o hotel
	in English: the wc is very uncomfortable for the hotel

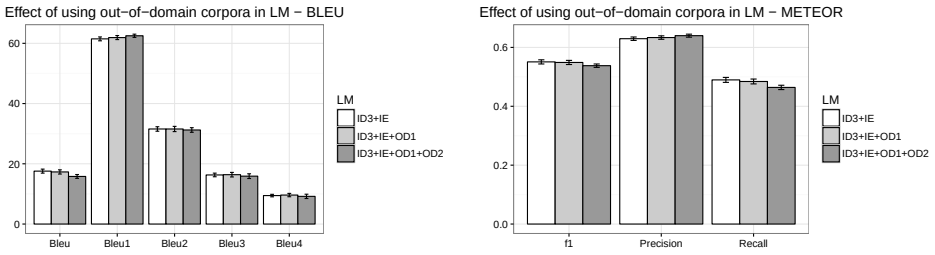

Figure 5 Effect in BLEU and METEOR scores of using additional out-of-domain corpora in the train of the Language Model. As baseline was used the system with the LM trained with ID3 plus IE corpora.

Figure 5 presents the results obtained by systems with three different language models. The first language model was obtained by output corpus plus in-domain corpus (as explained in previous section). On the second language model, we added, to previous LM, the OD1 corpus. Lastly, on the third language model, we added the OD1 and OD2 corpus.

The differences between all metrics (Bleu, Bleu1 to Bleu4) are not statistically significant. Just, a slight improvement is observed at Bleu1 metric, for system ID3+IE+OD1+OD2. The Meteor evaluation reveals that there are no significant differences from what is expressed at BLEU evaluation. However, is possible to observe that the addition of a out-of-domain corpus to language model increases the precision, with consequent reduction of recall values, meaning that translated sentences, by ID3+IE+OD1+OD2 system, are closer to the reference sentences, than sentences translated with system ID3+IE, which has simpler language models. Nevertheless, the improvement achieved is quite small, if we consider the dimension of OD1+OD2 corpora. One possible reason is the level of specialisation of ID3 corpus. Other, is its dimension, where, possibly, there are not enough sentences to learn the language model to build a large diversity of sentences.

5.2.4 Comparison with similar systems

Previous sections presented the evaluation results obtained by different variations of our system. Results obtained, by automatic evaluations, are consistent. Better results were obtained by ID3 corpus with variations of language models. Figure 6 compares the results from our best systems, with Mountain's best systems [4, pag. 73–75] and, for the best of our knowledge, the most recent system developed for data-to-text in Portuguese – MEDICATION2PT [7].

In this comparison, we used the system obtained from ID3 corpus, with two variations of language model: language model made only with the output language, and language model

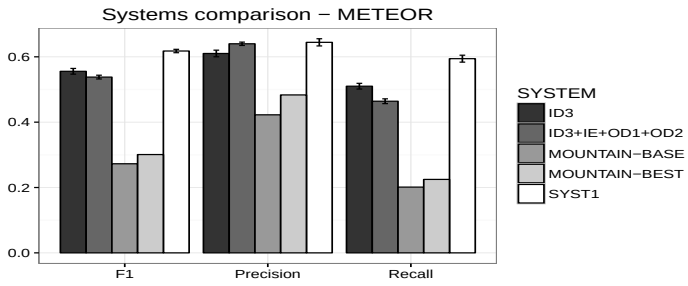

Figure 6 Comparison of Meteor scores of the best systems presented in this paper versus MOUTAIN system and a recent system developed for Portuguese (MEDICATION2PT).

made with output language plus IE and OD1 and OD2 corpora. From Mountain, we choose the original system and the system with higher results. From MEDICATION2PT, we choose the phrase-based version. By simplicity, only Meteor evaluations are presented, since BLEU results are similar. In all metrics, our systems' results are better than MOUNTAINS systems. At F_1 and Recall, results are almost twice the results obtained by MOUNTAIN systems.

When comparing with MEDICATION2PT (SYST1 in the Figure 6), the new system presents similar Precision but lower Recall and, consequently, lower F_1. The lower Recall is not necessarily negative, as we also aim at having variability in the generated sentences.

6 Conclusions

Aiming at providing a natural form (written text in Portuguese) of transmission of the information extracted automatically from comments on Hotels to managers, this paper presents a Data-to-Text aimed at this specific domain. The system is based in phrase-based machine translation performed by MOSES trained with small corpora. It is capable of processing information vectors produced automatically by an IE system to create sentences.

Several variants of the system were created to investigate several factors: method of corpus expansion; corpora used in training language models for output language, using both in-domain and out-of-domain corpora.

Using the automatic metrics BLEU and Meteor, the system variants were evaluated and compared with previous results. Results show that developed system variants are capable of achieving good Precision and producing usable sentences. A factor that showed significant effect in performance increase was corpus expansion. Results were significantly better when the base corpus was *expanded*. Use of additional corpora in language model training did not reveal capable of similar improvements in evaluation scores. But, as automatic metrics do not give a final answer regarding quality of the generated sentences, an evaluation by humans is needed to investigate further this factor.

Future work should also include exploration of mechanisms to combine generated sentences to produce texts.

Acknowledgements. The authors thank the reviewers for their contribution.

References

1 Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *6th Workshop on Statistical Machine Translation – EMNLP*, pages 85–91, 2011.

2 Flávio Ferreira, Nuno Almeida, Ana Filipa Rosa, André Oliveira, José Casimiro Pereira, Samuel Silva, and António Teixeira. Elderly centered design for interaction – the case of the S4S medication assistant. *Procedia Computer Science*, 27(Dsai 2013):398–408, 2014.

3 Albert Gatt and Ehud Reiter. SimpleNLG: a realisation engine for practical applications. In *12th European Workshop on Natural Language Generation*, pages 90–93, 2009.

4 Brian Langner. *Data-driven Natural Language Generation: Making Machines Talk Like Humans Using Natural Corpora.* PhD thesis, Carnegie Mellon University, 2010.

5 Eder Miranda Novais, Rafael Lage Oliveira, Daniel Bastos Pereira, Thiago Dias Tadeu, and Ivandre Paraboni. A testbed for portuguese natural language generation. In *Brazilian Symposium in Information and Human Language Technology*, pages 154–157, 2009.

6 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

7 José Casimiro Pereira and António Teixeira. Geração de linguagem natural para conversão de dados em texto – aplicação a um assistente de medicação para o português. *Linguamática*, 7(1):3–21, 2015.

8 José Casimiro. Pereira, António Teixeira, and Joaquim Sousa Pinto. Towards a hybrid NLG system for Data2Text in Portuguese. In *Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI)*, pages 679–684, 2015.

9 François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artifitial Intelligence*, 173(7-8):789–816, 2009.

10 Ehud Reiter. An architecture for data-to-text systems. In *Eleventh European Workshop on Natural Language Generation (ENLG)*, pages 97–104, 2007.

11 Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems.* Cambridge University Press, 2000.

12 Mário Rodrigues and António Teixeira. *Advanced Information Extraction.* Springer, 2015.

13 Diana Santos and Paulo Rocha. Evaluating CETEMPúblico, a free resource for portuguese. In *Annual Meeting of the Association for Computational Linguistics*, pages 442–449, 2001.

14 Douglas Fernandes Pereira Silva Junior, Ivandré Paraboni, and Eder Miranda Novais. Um sistema de realização superficial para geração de textos em Português. *RITA – Revista de Informática Teórica e Aplicada*, 20(3):31–48, 2013.

15 Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian P. Davy. Generating spatio-temporal descriptions in pollen forecasts. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, 2006.