PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 97 APRIL 2012 5–22

Towards Optimal Choice Selection for Improved Hybrid Machine Translation

Christian Federmann[a], Maite Melero[b], Pavel Pecina[d], Josef van Genabith[c]

[a] DFKI, German Research Center for Artificial Intelligence, Germany
[b] Barcelona Media, Spain
[c] CNGL, Dublin City University, Ireland
[d] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

Abstract

In recent years, machine translation (MT) research focused on investigating how hybrid MT as well as MT combination systems can be designed so that the resulting translations give an improvement over the individual translations.

As a first step towards achieving this objective we have developed a parallel corpus with source data and the output of a number of MT systems, annotated with metadata information, capturing aspects of the translation process performed by the different MT systems.

As a second step, we have organised a shared task in which participants were requested to build Hybrid/System Combination systems using the annotated corpus as input. The main focus of the shared task is trying to answer the following question: *Can Hybrid MT algorithms or System Combination techniques benefit from the extra information (linguistically motivated, decoding and runtime) from the different systems involved?*

In this paper, we describe the annotated corpus we have created. We provide an overview on the participating systems from the shared task as well as a discussion of the results.

1. Introduction

Machine translation is an active field of research with many competing paradigms to tackle core translation problems. In recent years, an important focus for research has been investigating how hybrid machine translation engines as well as combination systems including several translation engines can be designed and implemented so that the resulting translations give an improvement over the component parts.

Corresponding author: cfedermann@dfki.de
Cite as: Christian Federmann, Maite Melero, Pavel Pecina, Josef van Genabith. Towards Optimal Choice Selection for Improved Hybrid Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 97, 2012, pp. 5–22. doi: 10.2478/v10108-012-0001-1.

One of the main objectives in our research within the T4ME project[1] is to provide a systematic investigation and exploration of optimal choices in Hybrid Machine Translation supporting Hybrid MT design using sophisticated machine-learning technologies. As a first step towards achieving this objective we have developed a parallel corpus with source data and the output of a number of MT systems, representing carefully selected MT paradigms, annotated with metadata information, capturing aspects of the translation process performed by the different machine translation systems. Including detailed and heterogeneous system specific information as metadata in the translation output (rather than just providing strings) is intended to provide rich features for machine learning methods to optimise combination in hybrid machine translation.

This first version of the corpus is available online under `www.dfki.de/ml4hmt/` and comprises annotated outputs of five machine translation systems, namely Joshua, Lucy, Metis, Apertium, and the Moses-based PB-SMT component of MaTrEx. The language pairs supported by the corpus are: English↔German, English↔Spanish, English↔Czech (all in both directions).

In this paper, we describe the annotated corpus we have created—including the data used to obtain the sample corpus (Section 2), the translation engines applied when building the corpus (Section 3), and the format of the corpus (Section 4). We provide an overview on the challenge (Section 5) and give descriptions of the participating systems from the shared task (Section 6). This includes a comparison to a state-of-the-art system combination system. Using automated metric scores and results from a manual evaluation, we discuss the performance of the various combination systems and their implementations (Section 7). One interesting result from the shared task is the fact that we observed different systems winning according to the automated metrics and according to the manual evaluation. We conclude by summarising our research results and outline future work (Section 8).

2. Data

2.1. Annotation Data

As a source of the data to be included and annotated in the corpus we decided to use the WMT 2008 news test set, which is a set of 2,051 sentences from the news domain translated to the languages of interest (English, Spanish, German, Czech) and also some others (French, Hungarian). This test set was provided by the organizers of the Third Workshop on Machine Translation (WMT) in 2008 as test data for the shared translation task.

[1]EU FP7 funded Network of Excellence, grant agreement no.: 249119.

2.2. Training Data-Driven Systems

Some of the MT systems used in this work are data-driven (Joshua and MaTrEx). They require parallel data for translation phrase pair extraction, monolingual data for language modeling, and parallel development data for tuning of system parameters. Originally we intended to use the Europarl corpus (Koehn, 2005) for training purposes, but since the version of this widely used parallel corpus available at the time when the research reported in this article was carried out, did not include Czech, we used the Acquis (Steinberger et al., 2006) and News Commentary parallel corpora instead.

2.3. JRC-Acquis Multilingual Parallel Corpus

The JRC-Acquis Multilingual Parallel Corpus is an "approximation" of the Acquis Communautaire, the total body of European Union (EU) law applicable in the the EU Member States. It comprises documents that were available in at least in ten of the twenty EU-25 languages (official languages in the EU before 2007) and that additionally existed in at least three of the nine languages that became official languages with the EU Enlargement in 2004 (i.e. Czech, Hungarian, Slovak, etc.).

2.4. WMT News Commentary Parallel Corpus

The WMT News Commentary Parallel Corpus contains news and commentaries from the Project Syndicate and is provided as training data for the series of WMT translation shared tasks (`www.statmt.org`). Version 10 was released in 2010 and is available in English, French, Spanish, German, and Czech.

2.5. Development Data

The development data sets were taken from the WMT 2008 development data package. We chose the `nc-test2007` files, which consist of 2,007 sentences from the news-commentary domain available in English, French, Spanish, German, and Czech. These development sets not overlap with the training set.

3. System and Metadata Descriptions

3.1. Joshua

Description Joshua (Li et al., 2009), referred to as system t1 in the annotated corpus, is an open-source toolkit for statistical machine translation, providing a full implementation of state-of-the-art techniques making use of synchronous context free grammars (SCFGs). The decoding process features algorithms such as chart-parsing,

n-gram language model integration, beam-and cube-pruning and k-best extraction, while training includes suffix-array grammar extraction and minimum error rate training.

Annotation In our metadata annotations, we provide the output of the decoding process given the "test set", as processed by Joshua (SVN revision 1778). The annotation set contains the globally applied feature weights and for each translated sentence: the full output of the produced translation with the highest total score (among the n-best candidates), the language model and translation table scores, the scores from the derivation of the sentence (phrase scores) and merging/pruning statistics of the search process. Each translated sentence, represented by a hierarchical phrase, contains zero or more tokens and points to zero or more child phrases. Finally, the word-alignment of each phrase to the source text, using word indices, is available.

3.2. Lucy

Description The Lucy RBMT system (Alonso and Thurmair, 2003), system t2, uses a sophisticated RBMT transfer approach with a long research history. It employs a complex lexicon database and grammars to transform a source into a target language representation. The translation of a sentence is carried out in three major phases: analysis, transfer, and generation.

Annotation In addition to the translated target text Lucy provides information about the tree structures that have been created in the three translation phases and which have been used to generate the final translation of the source text. Inside these trees, information about POS, phrases, word lemma information, and word/phrase alignment can be found. In our metadata annotations, we provide a "flattened" representation of the trees. For each token, annotation may contain allomorphs, canonical representations, linguistic categories, or surface string.

3.3. Metis

Description The Metis system, system t3, (Vandeghinste et al., 2008) achieves corpus-based translation on the basis of a monolingual target corpus and a bilingual dictionary only. The bilingual dictionary functions as a flat translation model that provides n translations for each source word. The most probable translation given the context is then selected by consulting the statistical models built from the target language corpus.

Annotation Meta-data information for Metis is extracted from the set of final translations ranked by the Metis search engine. For each translation we obtain the score

computed during the search process, together with some linguistic information. The basic linguistic information provided is: lemma, POS tag, and morphological features, including gender, number, tense, etc.

3.4. Apertium

Description Apertium (Ramírez-Sánchez et al., 2006), system t4, originated as one of the machine translation engines in the project OpenTrad, funded by the Spanish government. Apertium is a shallow-transfer machine translation system, which uses finite state transducers for all of its lexical transformations, and hidden Markov models for POS tagging or word category disambiguation. Constraint Grammar taggers are also used for some language pairs (e.g., Breton-French).

Annotation We use Apertium version 3.2. Our metadata annotation includes tags, lemmas and syntactic information. We have used the following commands (in English-to-Spanish): en-es-chunker (for syntax information), en-es-postchunk (for tags and lemmas) and en-es (for the translation).

3.5. MaTrEx

Description The MaTrEx machine translation system (Penkale et al., 2010), system t5, is a combination-based multi-engine architecture developed at Dublin City University exploiting aspects of both the Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT) paradigms. The architecture includes various individual systems: phrase-based, example-based, hierarchical phrase-based, and tree-based MT. For the corpus data produced here we use the standard MOSES PB-SMT system (Koehn et al., 2007) as integrated into MaTrEx.

Annotation Sentence translations provided by MaTrEx in this work were obtained using the MOSES PB-SMT system decomposing the source side to phrases (n-grams), finding their translation and composing them to a target language sentence which has the highest score according the PB-SMT model. Meta-data annotations for each sentence translated by MaTrEx include scores from each model and is decomposed into phrases each provided with two additional scores: translation probability and future cost estimate. Additionally information about unknown words is also included.

4. Corpus Description

We have developed a new dedicated format derived from XLIFF (XML Localisation Interchange File Format) to represent and store the corpus data. XLIFF is an XML-based format created to standardize localization. It was standardized by OASIS in

2002 and its current specification is v1.2 (docs.oasis-open.org/xliff/xliff-core/xliff-core.html).

An XLIFF document is composed of one or more <file> elements, each corresponding to an original file or source. Each <file> element contains the source of the data to be localized and the corresponding localized (translated) data for one locale only. The localizable texts are stored in <trans-unit> elements each having a <source> element to store the source text and a <target> (not mandatory) element to store the translation.

We introduced new elements into the basic XLIFF format (in the "metanet" namespace) supporting a wide variety of metadata annotation of the translated texts by different MT systems (tools). The tool information is included in the <tool> element appearing in the header of the file. Each tool can have several parameters (model weights) which are described in <metanet:weight>.

Annotation is stored in <alt-trans> element within the <trans-unit> elements. The <source> and <target> elements in the <trans-unit> elements refer to the source sentence and its reference translation, respectively. The <source> and <target> elements in the <alt-trans> elements specify the input and output of a particular MT system (tool). Tool-specific scores assigned to the translated sentence are listed in the <metanet:scores> element and the derivation of the translation is specified in the <metanet:derivation> element. Its content is tool-specific.

The full format specification is available as an XML schema. An example annotation is depicted in Figure 1 at the end of this article.

4.1. Oracle Scores

At first, we compare the performance of the contributing systems (t1–t5) on the sentence level, using two popular metrics, 1-WER and smoothed-BLEU (Lin and Och, 2004). For this initial experiment, we worked on the language pair Spanish→English as this is also used in the ML4HMT shared task.

Table 1 shows the percentage of the cases that a system gave the best translation for a sentence according the two sentence-level metrics in columns 2 and 3.[2] Column 4 shows the overall BLEU score for the individual systems. This indicates that the systems included in the corpus perform complementary to each other.

In table 2 we show what the optimal BLEU performance would be, if a combination system was able to choose the best sentence of each component system according to each of the two metrics. This indicates the possibilities of improvement by a sentence-selection approach given the corpus. We believe that even higher performance would be possible using more sophisticated system combination methods.

[2]Measured over the development set, ties were allowed.

ranked 1st	1-WER[%]	sBLEU[%]	BLEU
system t1	26.44	14.73	12.80
system t2	37.56	38.63	14.94
system t3	5.85	5.85	8.29
system t4	16.00	23.41	13.34
system t5	58.54	29.95	14.47

Table 1. Preliminary investigation for the usability of the corpus for Hybrid MT

Oracle combination	BLEU
sBLEU-based	18.95
1-WER-based	17.62

Table 2. Potential BLEU score reachable using perfect sentence selection.

5. Challenge Description

The "Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT" is an effort to trigger systematic investigation on improving state-of-the-art Hybrid MT, using advanced machine-learning (ML) methodologies. Participants of the challenge are requested to build Hybrid/System Combination systems by combining the output of several MT systems of different types and with very heterogeneous types of metadata information, as provided by the organizers. The main focus of the shared task is trying to answer the following question:

> *Can Hybrid Machine Translation algorithms or System Combination techniques benefit from extra information—such as linguistic or linguistically motivated features, decoding parameters, or runtime annotation—from the different systems involved?*

The participants are given a bilingual development set, aligned at a sentence level. For each sentence, the corresponding *bilingual data set* contains:

- the source sentence,
- the target (reference) sentence, and
- the corresponding multiple output translations, annotated with metadata information, from five different systems, based on various machine translation approaches.

5.1. Development and Test Sets

We decided to use the WMT 2008 (Callison-Burch et al., 2008) news test set as a source for the annotated corpus.[3] This is a set of 2,051 sentences from the news domain translated to several languages, including English and Spanish but also others. The data was provided by the organizers of the Third Workshop on Machine Translation (WMT) in 2008. This data set was split into our own Hybrid MT Shared Task development set (containing 1,025 sentence pairs) and test set (containing 1,026 sentence pairs).

6. Combination Systems Participating in the Shared Task

6.1. DCU

The system described in Okita and van Genabith (2011) presents a system combination module in the MT system MaTrEx (Machine Translation using Examples) developed at Dublin City University. A system combination module deployed by them achieved an improvement of 2.16 BLEU (Papineni et al., 2001) points absolute and 9.2% relative compared to the best single system, which did not use any external language resources. The DCU system is based on system combination techniques which use a confusion network on top of a Minimum Bayes Risk (MBR) decoder (Kumar and Byrne, 2002).

One interesting, novel point in their submission is that for the given single best translation outputs, they tried to identify which inputs they will consider for the system combination, possibly discarding the worst performing system(s) from the combination input. As a result of this selection process, their BLEU score, from the combination of the four single best individual systems, achieved 0.48 BLEU points absolute higher than the combination of the five single best systems.

6.2. DFKI-A

A system combination approach with a sentence ranking component is presented in Avramidis (2011). The paper reports on a pilot study that takes advantage of multilateral system-specific metadata provided as part of the shared task. The proposed solution offers a machine learning approach, resulting in a selection mechanism able to learn and rank systems' translation output on the complete sentence level, based on their respective quality.

For training, due to the lack of human annotations, word-level Levenshtein distance has been used as a (minimal) quality indicator, whereas a rich set of sentence

[3] We deliberately did not use the WMT 2009 (Callison-Burch et al., 2009) news test set as there had been quality issues with this data set during the 2009 shared task.

features was extracted and selected from the dataset. Three classification algorithms (Naive Bayes, SVM and Linear Regression) were trained and tested on pairwise featured sentence comparisons. The approaches yielded high correlation with original, Levenshtein-based rankings ($\tau = 0.52$) and selected the best translation on up to 54% of the cases.

6.3. DFKI-B

The authors of Federmann et al. (2011) report on experiments that are focused on word substitution using syntactic knowledge. From the data provided by the workshop organisers, they choose one system to provide the "translation backbone". The Lucy MT system was suited best for this task, as it offers parse trees of both the source and target side, which allows the authors to identify interesting phrases, such as noun phrases, in the source and replace them in the target language output. The remaining four systems are mined for alternative translations on the word level that are potentially substituted into the aforementioned template translation if the system finds enough evidence that the candidate translation is better. Each of these substitution candidates is evaluated considering a number of factors:

- the part-of-speech of the original translation must match the candidate fragment.
- Additionally they may consider the 1-left and 1-right context.
- Besides the part-of-speech, all translations plus their context are scored with a language model trained on EuroParl.
- Additionally, the different systems may come up with the same translation, in that case the authors select the candidate with the highest count ("majority voting").

The authors reported improvements in terms of BLEU score when comparing to the translations from the Lucy RBMT system.

6.4. LIUM

Barrault and Lambert submitted results from applying the open-source MANY (Barrault, 2010) system on our data set. The MANY system can be decomposed into two main modules.

1. The first is the alignment module which actually is a modified version of TERp (Snover et al., 2009). Its role is to incrementally align the hypotheses against a backbone in order to create a confusion network. Each hypothesis acts as backbone, yielding the corresponding confusion network. Those confusion networks are then connected together to create a lattice.
2. The second module is the decoder. This decoder is based on the token pass algorithm and it accepts as input the lattice previously created. The costs computed

in the decoder can be expressed as a weighted sum of the logarithm of feature functions. The following features are considered in decoding:

- the language model probability, given by a 4-gram language model,
- a word penalty, which depends on the number of words in the hypothesis,
- a null-arc penalty, which depends on the number of null arcs crossed in the lattice to obtain the hypothesis, and
- the system weights: each word receives a weight corresponding to the sum of the weights of all systems which proposed it.

7. Evaluation Results

To evaluate the performance of the participating systems, we computed automated scores, namely BLEU, NIST, METEOR (Banerjee and Lavie, 2005), PER, Word error rate (WER) and Translation Error Rate (TER) and also performed an extensive, manual evaluation with 3 annotators ranking system combination results for a total of 904 sentences.

7.1. Automated Scores

The results from running automated scoring tools on the submitted translations are reported in Table 3. The overall best value for each of the scoring metrics is printed in **bold face**. The lower half of the table presents automated metric scores for the individual systems in the ML4HMT corpus, also computed on the test set. These scores give an indicative baseline for comparison with the system combination results. Again, the overall best value per column is printed in **bold face**. TER values are not available for the baseline systems as we initially did not intend to use this metric.

7.2. Manual Ranking

The manual evaluation is undertaken using the Appraise (Federmann, 2010) system; a screenshot of the evaluation interface is shown in Figure 2. Users are shown a reference sentence and the translation output from all four participating systems and have to decide on a ranking in *best-to-worst order*. Table 4 shows the average ranks per system from the manual evaluation, again the best value per column is printed in **bold face**. Table 5 gives the statistical mode per system which is the value that occurs most frequently in a data set.

7.3. Inter-annotator Agreement

Next to computing the average rank per system and the statistical mode, we follow Carletta (1996) and compute κ scores to estimate the inter-annotator agreement. In our manual evaluation campaign, we had $n = 3$ annotators so computing basic, pairwise

System	BLEU	NIST	METEOR	PER	WER	TER
DCU	**25.32**	**6.74**	56.82	**60.43**	**45.24**	**0.65**
DFKI-A	23.54	6.59	54.30	61.31	46.13	0.67
DFKI-B	23.36	6.31	**57.41**	65.22	50.09	0.70
LIUM	24.96	6.64	55.77	61.23	46.17	0.65
Joshua	19.68	6.39	50.22	47.31	62.37	–
Lucy	**23.37**	6.38	**57.32**	49.23	64.78	–
Metis	12.62	4.56	40.73	63.05	77.62	–
Apertium	22.30	6.21	55.45	50.21	64.91	–
MaTrEx	23.15	**6.71**	54.13	**45.19**	**60.66**	–

Table 3. Automated scores for participants and baseline systems on test set.

System	Annotator #1	Annotator #2	Annotator #3	Overall
DCU	2.44	2.61	2.51	2.52
DFKI-A	2.50	2.47	2.48	2.48
DFKI-B	**2.06**	**2.13**	**1.97**	**2.05**
LIUM	2.89	2.79	2.93	2.87

Table 4. Average rank per system per annotator from manual ranking of 904 (overlap=146) translations.

System	Ranked 1st	Ranked 2nd	Ranked 3rd	Ranked 4th	Mode
DCU	62	79	**97**	62	3rd
DFKI-A	73	65	**82**	80	3rd
DFKI-B	**127**	84	47	42	1st
LIUM	38	72	74	**116**	4th

Table 5. Statistical mode per system from manual ranking of 904 (overlap=146) translations.

Systems	π-Score	Systems	π-Score	Annotators	π-Score
DCU, DFKI-A	0.296	DCU, DFKI-B	0.352	#1,#2	0.331
DCU, LIUM	0.250	DFKI-A, DFKI-B	0.389	#1,#3	0.338
DFKI-A, LIUM	0.352	DFKI-B, LIUM	0.435	#2,#3	0.347

Table 6. Pairwise agreement (using Scott's π) for all pairs of systems/annotators. Note that scores in the last column are computed using all pairwise annotations available; these can be more than the overlapping N=146.

```
<trans-unit id="s71">
  <source xml:lang="es">El paciente fue aislado.</source>
    <target xml:lang="en">The patient was isolated.</target>
    <alt-trans rank="1" tool-id="t3">
      <source xml:lang="es">El paciente fue aislado.</source>
      <target xml:lang="en">The paciente was isolated .</target>
      <metanet:scores>
        <metanet:score type="total" value="-60.4375047559049"/>
      </metanet:scores>
      <metanet:derivation id="s71_t3_r1_d1">
        <metanet:phrase id="s71_t3_r1_d1_p1">
          <metanet:string>The</metanet:string>
          <metanet:annotation type="lemma" value="the"/>
          <metanet:annotation type="pos" value="AT0"/>
          <metanet:annotation type="morph_feat" value=":m:sg:"/>
          <metanet:alignment from="0" to="0"/>
        </metanet:phrase>
```

Figure 1. Example of annotation from the ML4HMT corpus.

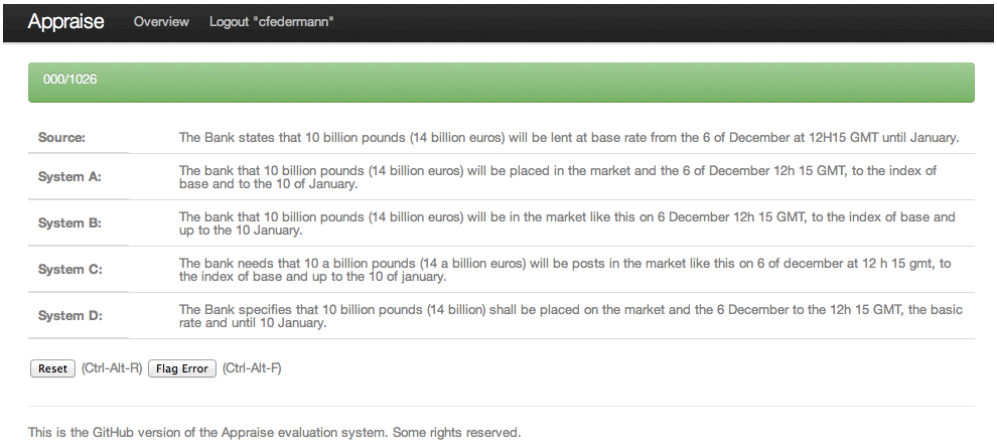

Figure 2. Screenshot of the Appraise interface for human evaluation.

annotator agreement is not sufficient—instead, we apply Fleiss (1971) who extends Scott (1955) for computing inter-annotator agreement for $n > 2$.

Annotation Setup As we have mentioned before, we had $n = 3$ annotators assign ranks to our four participating systems. As ties were not allowed, this means there exist $4! = 24$ possible rankings per sentence (e.g., *ABCD, ABDC, etc.*).[4] In a second evaluation scenario, we only collected the *1-best* ranking system per sentence, resulting in a total of four categories (A: "*system A ranked 1st*", etc.). In this second scenario, we can expect a higher annotator agreement due to the reduced number categories. Overall, we collected 904 sentences with an overlap of $N = 146$ sentences for which all annotators assigned ranks.

Scott's π allows to measure the pairwise annotator agreement for a classification task. It is defined as

$$\pi = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ represents the fraction of rankings on which the annotators agree, and $P(E)$ is the probability that they agree by chance. Table 6 lists the pairwise agreement of annotators for all four participating systems. Since we did not allow ties in the ranking process and because our ranks are not absolute categories, but can only be interpreted relatively to each other, we essentially have two categories for each pair of translations which are equally likely. Assuming $P(E) = 0.5$ we obtain an overall agreement π score of

$$\pi = \frac{0.673 - 0.5}{1 - 0.5} = 0.346 \tag{2}$$

which can be interpreted as *fair agreement* following Landis and Koch (1977). WMT shared tasks have shown this level of agreement is common for language pairs, where the performance of all systems is rather close to each other, which in our case is indicated by the small difference measured by automatic metrics on the test set (Table 3). The lack of ties, in this case might have meant an extra reason for disagreement, as annotators were forced to distinguish a quality difference which otherwise might have been annotated as "equal". We have decided to compute Scott's π scores to be comparable (or at least similar) to WMT11 (Bojar et al., 2011).[5]

Fleiss κ Next to the π scores, there also exists the so-called κ score. Its basic equation is strikingly similar to (1)

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3}$$

[4]Given this huge number of possible categories, we were already expecting resulting κ scores to be low.

[5]Note that we do not allow ties for the rankings and do not include the reference in the process, though.

with the main difference being the κ score's support for $n > 2$ annotators. We compute κ for two configurations. Both are based on $n = 3$ annotators and $N = 146$ sentences. They differ in the number of categories that a sentence can be assigned to (k).

1. *complete* scenario: $k = 24$ categories. For this, we obtained a κ score of

$$\kappa_{complete} = \frac{0.1 - 0.054}{1 - 0.054} = 0.049 \tag{4}$$

2. *1-best* scenario: $k = 4$ categories. Here, κ improved to

$$\kappa_{1-best} = \frac{0.368 - 0.302}{1 - 0.302} = 0.093 \tag{5}$$

It seems that the large number of categories of the *complete* scenario has indeed had an effect on the resulting $\kappa_{complete}$ score. This is a rather expected outcome, still we report the κ scores for future reference. The *1-best* scenario supports an improved κ_{1-best} score but does not reach the level of agreement observed for the π score.

It seems that DFKI-B was underestimated by BLEU scores, potentially due to its rule-based characteristics. This is a possible reason for the relatively higher inter-annotator agreement when compared with other systems. Also, DCU and LIUM may have low inter-annotator agreement as their background is similar. It is worth noting that METEOR was the only automated metric correlating with results from the manual evaluation.

Due to the fact that κ is not really defined for *ordinal data* (such as rankings in our case), we will investigate other measures for inter-annotator agreement. It might be a worthwhile idea to compute α scores, as described in Krippendorff (2004). Given the average rank information, statistical mode, π and κ scores, we still think that we have derived enough information from our manual evaluation to support for future discussion.

Cohen's κ As the results for Fleiss κ were disappointing for both settings, we also compute pairwise Cohen κ scores. Interestingly, we can again report *fair agreement* between the annotators, achieving κ scores similar to the π scores in table 6:

$$\kappa_{(\#1,\#2)} = 0.336 \qquad \kappa_{(\#1,\#3)} = 0.312 \qquad \kappa_{(\#1,\#3)} = 0.331 \tag{6}$$

The average Cohen's κ score is:

$$\kappa = 0.327 \tag{7}$$

8. Conclusion

We have developed an Annotated Hybrid Sample MT Corpus which is a set of 2,051 sentences translated by five different MT systems[6] (Joshua, Lucy, Metis, Apertium, and MaTrEx) in six translation directions (Czech→English, German→English, Spanish→English, English→Czech, English→German, and English→Spanish) and annotated with metadata information provided by the MT systems.

Using this resource we have launched the Shared Task on Applying Machine Learning techniques to optimise the division of labour in Hybrid MT (ML4HMT-2011), asking participants to create combined, hybrid translations using machine learning algorithms or other, novel ideas for making best use of the provided ML4HMT corpus data. The language pair for the shared task was Spanish→English.

Four participating combination systems, each following a different solution strategy, have been submitted to the shared task. We computed automated metric scores and conducted an extensive manual evaluation campaign to assess the quality of the hybrid translations. Interestingly, the system winning nearly all the automatic scores (DCU) only reached a third place in the manual evaluation. Vice versa, the winning system according to manual rankings (DFKI-B) ranked last place in the automatic metric scores based evaluation, with only one automated metric choosing it as winning system. This clearly indicates that more systematic investigation of hybrid system combination approaches, both on a system level and with regards to the evaluation of such systems' output, needs to be undertaken.

We have learnt from the participants that the metadata annotations provided by our ML4HMT corpus are possibly too heterogeneous to be used easily in system combination approaches; hence we will work on an updated version for the next edition of this shared task. Also, we will further focus on the integration of advanced machine learning techniques as these are expected to support better exploitation of our corpus' data properties.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119) and was supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) at Dublin City University and partially supported by the Czech Science Foundation (grant no. P103/12/G084). The authors would like to thank Felix Sasaki and Eleftherios Avramidis for their collaboration on the ML4HMT corpus and the shared task. Also, we are grateful to the anonymous reviewers for their valuable feedback and comments.

[6] Not all systems available for all language pairs.

Bibliography

Alonso, Juan A. and Gregor Thurmair. The Comprendium Translator System. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, 2003.

Avramidis, Eleftherios. DFKI System Combination with Sentence Ranking at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November 2011. META-NET.

Banerjee, Satanjeev and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0909`.

Barrault, Loïc. MANY : Open-Source Machine Translation System Combination. *Prague Bulletin of Mathematical Linguistics, Special Issue on Open Source Tools for Machine Translation*, 93: 147–155, 2010.

Bojar, Ondrej, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-2101`.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W08/W08-0309`.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March 2009. Association for Computational Linguistics.

Carletta, Jean. Assessing Agreement on Classification Tasks: the kappa Statistic. *Computational Linguistics*, 22:249–254, June 1996. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=230386.230390`.

Federmann, Christian. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valetta, Malta, May 2010. URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/197_Paper.pdf`.

Federmann, Christian, Yu Chen, Sabine Hunsicker, and Rui Wang. DFKI System Combination using Syntactic Information at ML4HMT-2011. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November 2011. META-NET.

Fleiss, J.L. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

Koehn, Philipp. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit 2005*, 2005.

Koehn, Philipp, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Annual meeting of the Association for Computation Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech, June 2007.

Krippendorff, Klaus. Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433, 2004.

Kumar, Shankar and William Byrne. Minimum Bayes-Risk Word Alignments of Bilingual Texts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 140–147, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1118693.1118712`.

Landis, J.R. and G.G. Koch. Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. URL `http://dx.doi.org/10.2307/2529310`.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. Joshua: An Open-Source Toolkit for Parsing-Based Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-0x24`.

Lin, Chin-Yew and Franz Josef Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1220355.1220427. URL `http://dx.doi.org/10.3115/1220355.1220427`.

Och, Franz Josef. Minimum Error Rate Training in Statistical Machine Translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA, 2003. Association for Computational Linguistics. URL `http://dx.doi.org/10.3115/1075096.1075117`.

Okita, Tsuyoshi and Josef van Genabith. DCU Confusion Network-based System Combination for ML4HMT. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011) and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation (ML4HMT)*, Barcelona, Spain, November 2011. META-NET.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176(W0109-022), IBM, 2001. URL `http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf`.

Penkale, Sergio, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. MaTrEx: the DCU MT system for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 143–148, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-71-8. URL `http://dl.acm.org/citation.cfm?id=1868850.1868870`.

Ramírez-Sánchez, Gema, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. OpenTrad Apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*, London, United Kingdom, November 2006. ISBN 0-85142-483-X.

Scott, William A. Reliability of Content Analysis: The Case of Nominal Scale Coding. *The Public Opinion Quarterly*, 19(3):321–325, 1955.

Snover, Matthew G., Nitin Madnani, Bonnie Dorr, and Richard Schwartz. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23:117–127, September 2009. ISSN 0922-6567. URL `http://dx.doi.org/10.1007/s10590-009-9062-9`.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, and Dan Tufis. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2142–2147, 2006.

Stolcke, Andreas. SRILM - An Extensible Language Modeling Toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado, 2002.

Vandeghinste, Vincent, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, Olga Yannoutsou, Toni Badia, Maite Melero, Gemma Boleda, Michael Carl, and Paul Schmidt. Evaluation of a Machine Translation System for Low Resource Languages: METIS-II. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Address for correspondence:
Christian Federmann
`cfedermann@dfki.de`

DFKI GmbH
Campus D3 2
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
GERMANY