ML4HMT Workshop Challenge at MT Summit XIII, Xiamen, China (http://www.dfki.de/ml4hmt/)

Challenge Description

The "Challenge on Optimising 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 are requested to build Hybrid/System Combination systems by combining the output of several systems of different types, which is provided by the organizers.

The main focus of the shared task is trying to answer the following question:

Could Hybrid/System Combination MT techniques benefit from extra information (linguistically motivated, decoding and runtime) from the different systems involved?

1) **Data:** The participants are given a development bilingual set, aligned at a sentence level. Each "bilingual sentence" contains:
   - the source sentence,
   - the target (reference) sentence and
   - the corresponding multiple output translations from 5 different systems, based on different MT approaches (*Apertium*, Ramírez-Sanchéz, 2006; *Joshua*, Zhifei Li et al, 2009; *Lucy*, Alonso and Thurmair, 2003; *Matrex*, Penkale et. al 2010) *Metis*, Vandeghinste et al., 2006). The output has been annotated with system-internal information deriving from the translation process of each of the systems (see below for a description of the data files, data format, and the employed MT systems.).
2) **Baseline:** As a baseline we consider state-of-the-art open-source system-combination systems, such as *MANY* (Barrault, 2010) and CMU-*MEMT* (Heafierld & Lavie, 2010).
3) **Challenge:** Participants are challenged to build an MT mechanism that improves over the baseline, by making effective use of the system-specific MT output. They can either provide solutions based on an open source system, or develop their own mechanisms. A suggested approach is given below.
   1. Spanish-English will be the language direction
   2. The development set can be used for tuning the systems during the development phase. Final submissions have to include translation output on a test set, which will be available one week before the submission deadline
   3. If you need language/reordering models they can be built upon the WMT News Commentary (http://www.statmt.org/wmt11/).
   4. Participants can also make use of additional linguistic analysis tools, if their systems require so, but they have to explicitly declare that upon submission, so that they are judged as "unconstrained" systems.
4) **Evaluation:** The system output will be judged via peer-based human evaluation. During the evaluation phase, participants will be requested to rank

system outputs of other participants through a web-based interface (*Appraise*; Federmann 2010). Automatic metrics (*BLEU*, Papineni et. al, 2002) will be additionally used.

5) **System description:** shared task participants will be invited to submit short papers (4-6 pages) describing their systems or their evaluation metrics (see instructions in **Submissions**).

Data Files

- news-test2008-dev.es-en.xml (development data)
- news-test2008-test.es-en.xml (test data)
- mt_in_xliff-for_meta-net.xsd, xliff_for_meta-net.xsd, xml.xsd (XML Schema files)

Annotated Data Format

We decided to use the WMT 2008 news test set as a source for the annotated corpus. 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 development set (containing 1025 sentence pairs) and test set (containing 1026 sentence pairs).

We have developed an own format to store the corpus data; it is derived from XLIFF (XML Localisation Interchange File Format) which is an XML-based format created to standardize localization. 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 (i.e. translated) data for one locale only. The localizable texts are stored in so-called `<trans-unit>` elements each having a `<source>` element to store the source text and a `<target>` element to store a translation, in our case both taken from the test set.

We introduced new elements into the basic XLIFF format (inside a dedicated "`metanet`" namespace) allowing annotation of the translated texts by different MT systems (tools). E.g., we store a tokenized version of the source text inside `<metanet:tokenized-source>`.

An example of a simple file in the XLIFF format is following:

```
<xliff version="1.2">

<file original="wmt2008-test" source-language="es" target-language="en">

<body>

<trans-unit id="s71">

<source xml:lang="es">El paciente fue aislado.</source>
```

```
<target xml:lang="en">The patient was isolated.</target>

<metanet:tokenized-source xml:lang="es">

<token>

<string>El</string>

</token>

<token>

<string>paciente</string>

</token>

...

<token>

<string>.</string>

</token>

</metanet:tokenized-source>

</trans-unit>

</body>

</file>

</xliff>
```

Example 1: XLIFF containing a sentence translated from Spanish to English.

Each tool can have several parameters (e.g. model weights), see Example 2.

```
...

<tool tool-id="t3" tool-name="Metis" tool-version="revision:2010">

<metanet:weights>

<metanet:weight type="lm" value="1.0"/>

<metanet:weight type="pt0" value="1.066893"/>

<metanet:weight type="pt1" value="0.752247"/>

<metanet:weight type="pt2" value="0.589793"/>

<metanet:weight type="wordpenalty" value="-2.844814"/>

</metanet:weights>

</tool>

...
```

Example 2: XLIFF extension describing the model weights for Metis.

Annotation of the translation is included in `<alt-trans>` element(s) 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 tokenized and lowercased input and output of a particular MT system (tool). Additionally, we store a true-cased and de-tokenized version of the translation inside `<detokenized-target>`. This represents the "final" translation which would e.g. be scored using BLEU. 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 and may contain part-of-speech annotations, alignment information or sub-phrases, etc. For illustration, see Example 3.

```
<trans-unit id="s71">

<source xml:lang="es">El paciente fue aislado.</source>

<target xml:lang="en">The patient was isolated.</target>

<metanet:tokenized-source>

...

</metanet:tokenized-source>

<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:detokenized-target>The paciente was isolated.</metanet:detokenized-target>

<metanet:scores>

<metanet:score type="total" value="-60.4375047559049"/>

</metanet:scores>

<metanet:derivation id="s71_t3_r1_d1">

...

</metanet:derivation>

</alt-trans>

</trans-unit>
```

Example 3: XLIFF extension describing the annotation of the translated text.

MT System descriptions

The data corpus we have prepared contains annotated translation outputs from five MT systems: Joshua, Lucy, Metis, Apertium, and Matrex for the Spanish–to–English direction.

In this section, we provide a brief description of the MT systems and details how they were used to create the annotated corpus.

Joshua

Joshua (system t1) is an open-source toolkit for statistical machine translation. It has been presented in 2009 by Johns Hopkins University, offering a full implementation of state-of-the-art techniques making use of *synchronous context free grammars* (SCFGs) (Zhifei Li et al. 2009). The decoding process features algorithms such as chart-parsing, *n*-gram language model integration, beam-and cube-pruning and *k*-best extraction, whereas training includes suffix-array grammar extraction and minimum error rate training.

Data

For the purposes of the annotated corpus, we performed a fresh training of Joshua-specific translation models, according to the specifications for the constrained statistical systems of the task: Acquis and WMT News Corpus were used for the extraction of the hierarchical grammar; the development set News Corpus dev-test 2007 was used for tuning the weights. As a result of the process, full decoder output is given for the WMT News-test 2008. All incoming corpora have been sentence-aligned (when parallel), tokenized and lowercased.

Training

The training process included *subsampling* of the training corpus, by keeping only the data that are required for the decoding of the tuning and the test corpus. *Berkeley Unsupervised Aligner* (ver. 2.1) was afterwards used in order to produce word alignments from the sentence-aligned aligned training corpora. The aligner made 5 iterations using two HMM alignment models trained jointly and then decoded using the competitive thresholding heuristic.

Given these alignments, Joshua scripts built a suffix-array and extracted the required SCFG grammar. An *n*-gram language model of order 5 was built over the target side of the corpus, by using the *SRILM toolkit* (ver. 1.5.11). Finally, the training process was concluded with a *Minimum Error Rate Training* with the use of *ZMERT script*, in order to get the model weights that optimize translation quality for the development set.

Recasing

As the lowercased output of the decoder requires *recasing*, we trained a SCFG grammar deriving recasing rules from the lowercased to the originally cased version of the training corpus. This grammar can be used to restore uppercase characters on the decoded output. This feature will be available in a future release of the annotation set.

Annotation

As a result of the annotation process, we provide the output of the decoding process given the "test set", as processed by Joshua (SVN revision 1778). The annotation set contains:

6) The globally applied model weights, as adjusted by the ZMERT process,
7) the full output of each translated sentence with the highest total score, among the *n*-best candidates,
8) the language model and translation table scores of each translated sentence
9) the derivation of each translated sentence The derivation is represented by defining hierarchical phrases in a tree-like structure;
10) each hierarchical phrase contains zero or more tokens and points to zero or more children phrases,
11) the word-alignment of each phrase to the source text, using word indices.

Joshua can provide an n-best hypotheses list, which will be integrated into a future release.

Lucy

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

1. **Analysis:** this phase takes place to identify the lexical units that are part of the translation unit (TU), and the way in which the TU is morpho-syntactically and semantically composed. It is by far the most complex phase during the translation. The output of the analysis phase is the analysis tree. This analysis tree can be regarded as a source language dependent representation of the meaning of the input TU.
2. **Transfer:** the next phase in the translation process is the transfer, during which the source language dependent semantic representation is transferred to

a target language dependent semantic representation. Thus, transfer is language-pair specific. During transfer, the Lucy system accesses the directional transfer lexicon. The result of the transfer is a transfer tree that contains target language words as terminal nodes; these words are still in their canonical form.

3. **Generation:** the purpose of the final phase, generation, is to produce the output text in the target language. The structural and lexical aspects of translation have already been treated by the transfer, so that generation only has to produce the surface representation of the target language TU. Surface creation has two aspects: allomorph choice and inflectional suffixing. Generation is language-pair independent, as it only requires access to the target language monolingual lexicon.

Next to the translated target text Lucy allows to export 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 part of speech, phrases, word lemma information, and word/phrase alignment can be found.

Annotation

As a result of the annotation process, we provide a "flattened" representation of the trees for the three translation phases of the Lucy MT engine. For each token annotation may contain allomorphs (ALO), canonical representations (CAN), linguistic categories (CAT), surface string (STRING). We plan to release a more refined representation of the Lucy trees in a future release o this corpus.

Metis

The Metis system (system t3, Vandeghinste et al., 2006) 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 off the TL corpus.

More specifically, the Spanish-English system uses only very basic linguistic resources to pre-process the input sentences, namely a POS tagger and lemmatiser, whose output is a string of lemmas or base forms, with disambiguated POS tags and inflectional information. Morphological disambiguation is performed by selecting the most plausible reading for each word given the context. At a subsequent step, morphological tags are mapped into the Parole/EAGLES tagset used by the bilingual dictionary. In this mapping step, information about POS, which will be used during dictionary look-

up, is separated from inflectional information which will be used only later, in token generation. Lexical translation is performed by a lemma-to-lemma dictionary, which contains information about the POS of both the source and the target word. No structure transfer rules are used. The output of the SL preprocessing and dictionary look-up is a set of translation candidates in form of strings of English lemmas and POS tags, ordered according to Spanish-like syntax. A series of target language models are built by indexing all the n-grams for $1 \leq n \leq 4$.

N-grams can belong to two different types:

- a sequence of lemma/tag (e.g. always/ADV + wear/VV + a/AT + hat/NN)
- a sequence of lemma/tag except for one position of tag alone (e.g. ADV + wear/VV + /AT + hat/NN)

During the indexing process, tokens are usually indexed as either lemma/tag or tag alone. Exceptions are:

- personal pronouns (PNP) which are always lemma/tag
- cardinals (CRD), ordinals (ORD) and unknown words (UNC) which are always indexed as tag alone.

To account for structure modifications, we allow permutation of CWs between two consecutive boundaries, as well as insertion and deletion of a predefined set of Function Words (prepositions, pronouns...).The decoder performs a beam search decoding over the n-gram models and outputs a ranked set of translations. At the final step, word form generation is performed, and some post-generation rules accounting for surface phenomena apply.

Annotation

The XML data gathered from Metis is extracted from the set of final translations ranked by the Metis search engine. For each translation we get the score computed during the search process, together with some linguistic information. The basic linguistic information provided is: lemma, part-of-speech tag, and morphological fratures. Morphological features are grouped under one feature and come from the source token. They may refer to gender, number, tense, etc.

Apertium

Apertium (system t4) originated as one of the machine translation engines in the project OpenTrad, which was funded by the Spanish government. It was originally designed to translate between closely related languages, although it has recently been expanded to treat more divergent language pairs. To create a new machine translation

system, one just has to develop linguistic data (dictionaries, rules) in well-specified XML formats.

Apertium is a shallow-transfer machine translation system, which uses finite state transducers for all of its lexical transformations, and hidden Markov models for part-of-speech tagging or word category disambiguation.

Annotation

We have use the stable version of Apertium (3.2) available at http://sourceforge.net/. The output includes tags, lemmas and syntactic information. We have used the following commands (in Spanish-to-English): es-en-chunker (for syntax information), es-en-postchunk (for tags and lemmas) and es-en (for the translation).

Matrex

The Matrex machine translation system (system t5) is a combination-based multi-engine architecture developed at Dublin City University (e.g. Penkale et. Al 2010) 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 first version of this deliverable, we only exploited the SMT phrase-based component of the system which is based on Moses (Koehn et. Al 2007) – an open-source toolkit for statistical machine translation which includes a wide variety of tools for training, tuning and decoding (applying the system).

Data

The Matrex system was trained in both translation directions for all language pairs (see Table 1). For training we exploited both parallel corpora available for the project: the Acquis corpus and WMT News Commentary corpus. Language models were trained on the target sides of the corpora. The WMT 2007 test set was used as the development set for parameter optimization.

Preprocessing and training

Prior training, all data were tokenized and lowercased using the standard Europarl tools (http://www.statmt.org/europarl/). The original (non-lowercased) versions of the target sides of the parallel data were kept for training the recasing language model. The lowercased versions of the target sides were used for language model training using the the SRILM toolkit (Stolcke 2002). Translation models were trained on the Acquis and WMT News Commentary corpora filtered on sentence level – we kept all sentence pairs having less than 100 words on each side and with length ratio within the interval

<0.9,9.0> which reduced the size of the corpus by 5% in average. Minimum error rate training (MERT, Och 2003) was employed to optimize the model parameters on the WMT 2007 test set.

Decoding and postprocessing

The annotation data WMT 2008 test set was lowercased, tokenized, and translated by the trained systems. Letter casing was reconstructed by the recasing model and extra spaces in the tokenized text were removed in order to produce correct and readable text.

Annotation

Sentence translations provided by Matrex in this work were obtained by 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 model. Thus, each sentence translated by Matrex is provided with scores from each model and decomposed to phrases each provided with two additional scores: translation probability and future cost estimate (for details, see Moses manual at http://www.statmt.org/moses/). Information about unknown words is also included.

About Data preparation and Contact

The annotated data corpus for this challenge has been prepared as part of the dissemination effort of META-NET (http://www.meta-net.eu/), a network of excellence dedicated to building the technological foundations of a multilingual European information society.

For more information about the challenge or the workshop, please contact us at: ml4hmt@easychair.org