Shared task

DiscoMT 2015 shared task on pronoun translation

Overview of the task

The DiscoMT 2015 shared task will consist of two subtasks, relevant to both the MT and discourse communities: **pronoun-focused translation**, a practical MT task, and **cross-lingual pronoun prediction**, a classification task that requires no specific MT expertise and is interesting as a machine learning task in its own right. For groups wishing to participate in both tasks, one possibility is to convert a system for the classification task into an MT feature model using existing software such as the Docent decoder (Hardmeier et al., ACL 2013). Both tasks use the English–French language pair, which has a sufficiently high baseline performance to produce basically intelligible output, as well as interesting differences in their pronoun systems.

Pronoun-Focused Translation Task

In the pronoun-focused translation task, you are given a collection of English input documents, which you are asked to translate into French. This task is the same as for other MT shared tasks such as that of WMT. The difference is in the way the translations are evaluated. Instead of checking the overall translation quality, we specifically look at how the English subject pronouns *it* and *they* were translated. The principal evaluation will be carried out manually and will focus specifically on the correctness of pronoun translation. Thanks to a grant from the EAMT, the manual evaluation will be run by the organisers and participants don't have to contribute evaluations. Automatic reference-based metrics are available for development purposes.

The texts in the test corpus will consist of transcripts of TED talks. The training data contains an in-domain corpus of TED talks as well as some additional data from Europarl and news texts. To make the participating systems as comparable as possible, we ask you to constrain the training data of your system to the resources listed below as far as you can, but this is not a strict requirement and we do accept submissions using additional resources. If your system uses any resources other than those of the official data release, please be specific about what was included in the system description paper. For the same reason, we also suggest that you use the tokeniser provided by us unless you have a good reason to do otherwise.

The test set will be supplied in the XML source format of the 2009 NIST MT evaluation, which is described on the last page of this document. See the development set included in the data release for an example. Your translation should be submitted in the XML translation format of the 2009 NIST MT evaluation. We also need you to submit, in a separate file, word alignments linking occurrences of the pronouns *it* and *they* (case-insensitive) to the corresponding words generated by your MT system. The format of the word alignments should be the same as that of the alignments included in the cross-lingual pronoun prediction data (see below). Word alignments can be obtained, for instance, by running the Moses SMT decoder with the -print-alignment-info option or by parsing the segment-level comments added to the output by the Docent decoder. You may submit alignments for the complete sentence if it's easier for you, but only links for *it* and *they* will be used. If your MT system cannot output word alignments, please contact the shared task organisers to discuss how to proceed. We'll try to find a solution. More details on how to submit will be added to this page later.

The test set will be released on **4 May 2015**, and your translations are due on **10 May 2015**. Detailed submission instructions can be found at the end of this page. Note that we will ensure that each document in the test set contains an adequate number of challenging pronouns, so the corpus-level distribution of the pronouns in the test set may differ from that of the training corpus. However, each document will be a complete TED talk with a naturally occurring ensemble of pronouns.

Cross-Lingual Pronoun Prediction

In the cross-lingual pronoun prediction task, you are given an English document with a human-generated French translation and a set of word alignments between the two languages. In the French translation, the words aligned to the English third-person subject pronouns *it* and *they* are substituted by placeholders. Your task is to predict, for each placeholder, the word that should go there from a small, closed set of classes, using any information you can extract from the documents. The following classes exist:

ce	The French pronoun *ce* (sometimes with elided vowel as *c'*) as in the expression *c'est* 'it is'
elle	feminine singular subject pronoun
elles	feminine plural subject pronoun
il	masculine singular subject pronoun
ils	masculine plural subject pronoun
ça	demonstrative pronoun (including the misspelling *ca* and the rare elided form *ç'*)
cela	demonstrative pronoun
on	indefinite pronoun
OTHER	some other word, or nothing at all, should be inserted

This task will be evaluated automatically by matching the predictions against the words found in the reference translation by computing the overall accuracy and precision, recall and F-score for each class. The primary score for the evaluation is the macro-averaged F-score over all classes. Compared to accuracy, the macro-averaged F-score favours systems that consistently perform well on all classes and penalises systems that maximise the performance on frequent classes while sacrificing infrequent ones.

The data supplied for the classification task consists of parallel English-French text with word alignments. In the French text, a subset of the words aligned to English occurrences of *it* and *they* have been replaced by placeholders of the form *REPLACE_xx*, where *xx* is the index of the English word the placeholder is aligned to. Your task is to predict one of the classes listed above for each occurrence of a placeholder.

The training and development data is supplied in a file format with five tab-separated columns:

1. the class label
2. the word actually removed from the text (may be different from the class label for class OTHER and in some edge cases)
3. the English source segment
4. the French target segment with pronoun placeholders
5. the word alignment (a space-separated list of alignments of the form SRC-TGT, where SRC and TGT are zero-based word indices in the source and target segment, respectively)

A single segment may contain more than one placeholder. In that case, columns 1 and 2 contain multiple space-separated entries in the order of placeholder occurrence. A document segmentation of the data is provided in separate files for each corpus. These files contain one line per segment, but the precise format varies depending on the type of document markup available for the different corpora. In the development and test data, the files have a single column containing the ID of the document the segment is part of.

Here's an example line from one of the training data files:

```
elles	Elles	They arrive first .	REPLACE_0 arrivent en premier .	0-0 1-1 2-3 3-4
```

The test set will be supplied in the same format, but with columns 1 and 2 (*elles* and *Elles)* empty, so each line starts with two tab characters. Your submission should have the same format as column 1 above, so a correct solution would contain the class label *elles* in this case. Each line should contain as many space-separated class labels as there are REPLACE tags in the corresponding segment. For each segment not containing any REPLACE tags, an empty line should be emitted. Additional tab-separated columns may be present in the submission, but will be ignored. Note in particular that you are not required to predict the second column. The submitted files should be encoded in UTF-8 (like the data we provide).

The test set will be the same as for the pronoun-focused translation task. The complete test data for the classification task, including reference translations and word alignments, will be released on **11 May 2015**, after the completion of the translation task. Your submission is due on **18 May 2015**. Detailed submission instructions can be found at the end of this page.

Note: If you create a classifier for this task, but haven't got an MT system of your own, you might consider using your classifier as a feature function in the document-level SMT decoder Docent to create a submission for the pronoun translation task.

Discussion Group

If you are interested in participating in the shared task, we recommend that you sign up to our discussion group to make sure you don't miss any important information. Feel free to ask any questions you may have about the shared task!

https://groups.google.com/d/forum/discomt2015

Training Data and Tools

All training and development data for both subtasks can be downloaded from the following locations:

- https://www.dropbox.com/sh/c8qnpag5z29jyh6/AAAAqk1TE9-UvcgEnfccdRwxa?dl=0
- Download alternative 1: http://opus.lingfil.uu.se/DiscoMT2015/
- Download alternative 2: http://stp.lingfil.uu.se/~joerg/DiscoMT2015/

The dropbox folder contains many files, see the list below and the README file. To create a system for the pronoun classification task, you should start with the classification training data. For the pronoun-focused translation task, we provide both the original training data, preprocessed data sets including full word alignments and a complete pre-trained phrase-based SMT system. To minimise preprocessing differences among the submitted system we suggest (but do not require) that you start from the most processed version of the data that is usable for the type of system that you plan to build.

Classification training data including pre-processing scripts and doc-IDs:
DiscoMT2015.classification.tar.bz2
(the important files are in DiscoMT/classification/*.data.gz)

Raw training data for SMT (monolingual and parallel, including doc-IDs and markup where available):
DiscoMT2015.monolingual-data.tar.bz2
DiscoMT2015.parallel-data.tar.bz2

Tokenized and lowercased versions of the same:
DiscoMT2015.tokenized-monolingual-data.tar.bz2
DiscoMT2015.tokenized-parallel-data.tar.bz2

Europarl and TED data with XML markup:
DiscoMT2015.annotated-data.tar.bz2

More meta-data for TED:
DiscoMT2015.IWSLT14-info.tar.bz2

Devset with XML markup and document IDs (last file):
IWSLT14.TED.tst2010.en-fr.en.xml
IWSLT14.TED.tst2010.en-fr.fr.xml
TEDdev.en-fr.doc-ids

Baseline SMT model including pre-processing scripts used to prepare the above data sets:
DiscoMT2015.baseline.tar.bz2

Baseline SMT model including intermediate files (like word alignment etc):
DiscoMT2015.baseline-all.tar.bz2

Recaser
http://stp.lingfil.uu.se/~joerg/DiscoMT2015/DiscoMT2015.recaser.tar.gz

XML wrapper for MT output
http://stp.lingfil.uu.se/~ch/source-to-test.tar.gz

Evaluation

The classification results will be evaluated against the gold standard translations from the test set (see the example for the classification baseline below). The current version of the scorer is available here: discoMT_scorer.pl.

For the pronoun-focused translation task, the submissions will be scored manually. For your convenience, there is an implementation of the automatic pronoun precision/recall scorer by Hardmeier and Federico (IWSLT 2010) available on GitHub, but please not that this is NOT an official scorer and that it shouldn't be considered a reliable evaluation metric.

Classification Baseline

We have a baseline model for the classification task that looks only at the language model scores (using KenLM, which you can get here, and the language model that is used needs to be in KenLM's binary format (which is the case for the "corpus.5.trie.kenlm" included in the "baseline-all"

tarball).

You can get predictions for the baseline model by running, e.g.

python discomt_baseliny.py --fmt=replace .../classification/TEDdev.data.gz

These are in the format that the scorer requires, with predictions in the first column, the word it predicted in the second column (which is *always* ignored by the scorer, so don't worry if your system doesn't predict words), etc.

If you're interested in just using the marginal probabilities for each filler from the language model, you can also use

python discomt_baseline --fmt=scores .../classification/TEDdev.data.gz

which will give you, for each input line, one with TEXT in the second column giving you the source/target text, and zero or more lines with ITEM 0, ITEM 1 etc. giving you a (partial) probability distribution over the fillers for each "REPLACE" position.

Other flags:

- --lm LANGUAGE_MODEL: use another language model (otherwise it assumes the default name and the current directory)
- --null-penalty PENALTY: use this penalty for predicting no filler at all (which counts as OTHER)

Results with default options on TEDdev (same data as tst2010):

```
   ce : P = 110/ 129 = 85.27% R = 110/ 148 = 74.32% F1 = 79.42%
 cela : P = 4/ 15 = 26.67%    R = 4/ 10 = 40.00%    F1 = 32.00%
 elle : P = 6/ 13 = 46.15%    R = 6/ 30 = 20.00%    F1 = 27.91%
elles : P = 4/ 12 = 33.33%    R = 4/ 16 = 25.00%    F1 = 28.57%
   il : P = 35/ 137 = 25.55%  R = 35/ 55 = 63.64%   F1 = 36.46%
  ils : P = 86/ 94 = 91.49%   R = 86/ 139 = 61.87%  F1 = 73.82%
   on : P = 3/ 10 = 30.00%    R = 3/ 10 = 30.00%    F1 = 30.00%
   ça : P = 16/ 22 = 72.73%   R = 16/ 61 = 26.23%   F1 = 38.55%
OTHER : P = 225/ 315 = 71.43% R = 225/ 278 = 80.94% F1 = 75.89%
```

or a macro-averaged fine-grained F1 of 46.96%

Results with "--null-penalty -2.0"

```
   ce : P = 121/ 145 = 83.45% R = 121/ 148 = 81.76% F1 = 82.59%
 cela : P = 4/ 21 = 19.05%    R = 4/ 10 = 40.00%    F1 = 25.81%
 elle : P = 7/ 15 = 46.67%    R = 7/ 30 = 23.33%    F1 = 31.11%
elles : P = 5/ 14 = 35.71%    R = 5/ 16 = 31.25%    F1 = 33.33%
   il : P = 36/ 143 = 25.17%  R = 36/ 55 = 65.45%   F1 = 36.36%
  ils : P = 99/ 109 = 90.83%  R = 99/ 139 = 71.22%  F1 = 79.84%
   on : P = 3/ 13 = 23.08%    R = 3/ 10 = 30.00%    F1 = 26.09%
   ça : P = 19/ 32 = 59.38%   R = 19/ 61 = 31.15%   F1 = 40.86%
OTHER : P = 211/ 255 = 82.75% R = 211/ 278 = 75.90% F1 = 79.17%
```

or a fine-grained F1 score of 48.35%

Important dates

February 2015	Training data release
4 May 2015	Release of test data for pronoun-focused translation task
10 May 2015	Submission deadline for pronoun-focused translation task
11 May 2015	Release of test data for cross-lingual pronoun prediction task
18 May 2015	Submission deadline for cross-lingual pronoun prediction task
28 June 2015	System paper submission deadline
21 July 2015	Notification of acceptance
11 August 2015	Camera-ready papers due
September 2015	DiscoMT 2015 workshop in Lisbon (in conjunction with EMNLP)

Submission Instructions

For the **pronoun-focused translation task**, the test set will be released on 4 May 2015 as a NIST-XML file in raw text form and in tokenised, lowercased form to match the preprocessing of the training data and baseline system we provided. Your submission should consist of three files:

1. The *tokenised* version of the *input data* in NIST-XML (srcset) format. This may be identical to the tokenised input file we provide to you unless you used a different tokeniser.
2. The *tokenised and recased* version of your *machine translation output* in NIST-XML (tstset) format.
3. A file containing the word alignments of the input pronouns *it* and *they*. This file should be a plain text file (no XML markup) containing exactly as many lines as there are segments in the test set. For each segment, there should be a (possibly empty) line with the word alignment links (a space-separated list of alignments of the form SRC-TGT, where SRC and TGT are zero-based word indices in the source segment in file [1] and the MT output segment in file [2], respectively). The word alignment file may, but need not, contain word alignments for other words as well. These will be ignored in the evaluation.

Please wrap up these three files in a single *tar.gz* or *zip* file labelled with the name of your system and e-mail it to *discomt2015-submissions@rax.ch* no later than 10 May 2015 (any time zone).

To make sure that your submitted files conform to the instructions, we suggest that you validate them before submitting. For the XML files, you should at least check that they are well-formed. This can be done very easily with a tool like XML Starlet - just run *xml val file.xml*. We also provide a DTD that you can validate your files against. With XML Starlet, you'd run *xml val -d mteval-xml.DiscoMT2015.dtd file.xml* to do so. For the word alignment file, you should at least use *wc -l* to check that the number of lines is correct and equals the number of segments in the test set. There will be 2093 segments in the test set.

For the **cross-lingual pronoun prediction task**, we provide the input data in the same format as the training data, but with the first two columns empty. Your predictions should be submitted in the format recognised by the official scorer, see above for details. Please e-mail the file with the predictions, labelled with the name of your system, to *discomt2015-submissions@rax.ch* no later than 18 May 2015 (any time zone).

Test Data

The test data for the pronoun-focused translation task can be downloaded from the following locations:

	Source	Reference
raw text	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.raw.en.xml	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.ra
tokenised text (not lowercased)	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.tok.en.xml	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.to
tokenised and lowercased text	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.low.en.xml	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.lo

The test data for the cross-lingual pronoun prediction task can be found here:

classification test data (including answers)	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.gold.gz
document IDs	http://stp.lingfil.uu.se/~ch/DiscoMT2015.test/DiscoMT2015.test.doc-ids.gz

System Description Papers

All groups who participated in the DiscoMT shared task are invited to submit a paper to the workshop to describe their system in detail. System description papers can be up to 6 pages long (excluding references) and will be presented as posters at the workshop. They should be submitted through the "Shared task" track of the DiscoMT START system. System description papers are subject to review and may be rejected if the quality of the description is strongly insufficient. However, the scores or ranking achieved in the shared task evaluation have no bearing on the acceptance decision, and you are welcome to present your system at the workshop no matter how successful your approach was in the evaluation.

All system description papers should contain enough details to make the results reproducible. They need not provide a detailed description of the task itself and the data sets provided by the organisers. Instead, they may refer to the shared task overview paper, whose bibliographic details will be announced before the camera-ready deadline. Data sets and tools not included in the official data release should be described. We also welcome a critical discussion of the system performance and encourage including additional contrastive results. Unlike regular long and short papers, system description papers need not be anonymised. The paper submission deadline is 28 June 2015.

Evaluation Results

The manual evaluation of the pronoun-focused translation task is ongoing (as of 1 June). We expect to publish the results by 23 June. Automatic metrics for the translation task will be published at the same time.

Initial results for the pronoun prediction task are already available here. More details will be added to that page before mid-June.

Acknowledgements

Funding for the manual evaluation of the pronoun-focused translation task is generously provided by the European Association for Machine Translation.