EVALUATION PLAN FOR SPANISH-BASQUE MACHINE TRANSLATION

1.- Introduction

The purpose of this evaluation is to promote the research in machine translation between Spanish and Basque, and to foster collaboration between groups, since there are fewer resources, both bilingual and monolingual, related to Basque.

With this aim, we propose an evaluation plan to translate informative consumption articles from Spanish to Basque. We will accept two approaches. The first is to develop a translator using all the teams the same resources, to be able to compare different approaches under the same conditions. The second option will allow to present systems using other available resources, such as additional corpus, different language processors or dictionaries.

Participants are committed to presenting the results of the evaluation at a special session to be held during the V Biennial Workshop on Speech Technology. Participation may be individual or in a team where the representative must be a student. Each team may submit one or more systems.

2.- Measuring performance

The performance will be measured using several automatic scoring techniques. These measures will compare the output of machine translation with the reference manual translation. We will apply the usual measures BLEU, NIST, WER and PER. BLEU[1] and NIST[2] measures are based on the similarity of subsequences (N-grams) of the machine translation and the reference. The word error rate WER measures insertions, deletions and substitutions between the machine translation and the reference. The position independent word error rate PER calculates the distance between the set of words of the machine translation and the one of the reference. For the evaluation, we will use the tools from TC-STAR (www.tc-star.org).

3.- Terms of evaluation

We will assess two types of systems: systems based on the provided resources and systems without restrictions. For systems based on limited resources, we will provide 58,000 sentences for training, approximately 1,500 to adjust the system and another 1,500 to test. There will be a single translation for each sentence in the source.

The corpus will be provided in text format, tokenized and aligned at sentence level, and the corpus corresponding to each language should be delivered separately. Alternatively, we should provide the text lemmatized and tagged automatically. Each line in the file will correspond to a sentence for both source and target. All the processes used are automatic, so the corpus will contain errors. The final test will be to translate several articles of the same domain but not included in the initial material, and will be processed in the same way.

---

[1]K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proc. of the 40th Annual Meeting of the ACL, Philadelphia, PA, July 2002, pp. 311–318.

[2]Doddington, G. "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics". In Proc. ARPA Workshop on Human Language Technology, San Diego, California, March 2002.

4.- Procedure for evaluation

The schedule for the evaluation is the following:

- On May 1st, 2008 the evaluation plans will be announced and the registration period will begin.
- The deadline for registration will be on May 31st, 2008.
- After June 16th, 2008 the training and development material for the evaluations will be available. It is a must to be registered for the evaluation to receive the material.
- On September 15th, 2008 the test data will be provided.
- September 30th, 2008 at 24:00 is the deadline for receiving the results.
- On October 31st, 2008 the results of the evaluation will be published among participants.

5.- Sending translations

The translations will be sent by mail to the organization. They must be complete, thus containing the entire dataset for evaluation. They should be sent to: Nerea Ezeiza (nerea.ezeiza@gmail.com). A file per system and per article to be evaluated must be sent.

All the results will be available once they have been sent to each participant. This will allow participant analyse them prior to the V Workshop on Speech Technologies.

Each participant must send a description of the system sent for the evaluation, which should include:

- Identity chosen for the system (sysid)
- Conditions of evaluation (training data)
- In the case of systems without restrictions, you should indicate the characteristics of the resources used for development:
  - In the case of the corpus you should indicate:
    - if the provided corpus has been used or not
    - if an alternative corpus has been used, a brief description of the contents, approximate size, if it is monolingual or bilingual and if it is publicly available
  - In the case of dictionaries, approximate size, whether they are monolingual or bilingual and if there are publicly available
  - In the case of language processors:
    - processor type (morphological analyzer / POS tagger / shallow syntactic parser / semantic processor /...)
    - if it is publicly available
    - languages to which they have been applied
- description of the algorithmic approach

This description will be sent using the template used for the regular communications of the V Workshop on Speech Technology. The descriptions received will be distributed as part of the analytic material of evaluation.