Shared Task: Statistical Machine Translation between European Languages

Philipp Koehn
School of Informatics
University of Edinburgh
Edinburgh, EH8 9LW, UK
pkoehn@inf.ed.ac.uk

Christof Monz
UMIACS
University of Maryland
College Park, MD 20742, USA
christof@umiacs.umd.edu

Abstract

The ACL-2005 Workshop on Parallel Texts hosted a shared task on building statistical machine translation systems for four European language pairs: French–English, German–English, Spanish–English, and Finnish–English. Eleven groups participated in the event. This paper describes the goals, the task definition and resources, as well as results and some analysis.

Statistical machine translation is currently the dominant paradigm in machine translation research. Annual competitions are held for Chinese–English and Arabic–English by NIST (sponsored by the US military funding agency DARPA), which creates a forum to present and compare novel ideas and leads to steady progress in the field.

One of the advantages of statistical machine translation is that the currently applied methods are fairly language-independent. Building a new machine translation system for a new language pair is not much more than a matter of running a training process on a training corpus of parallel text (a text in one language paired with a translation in another).

It is therefore possible to hold a competition where research groups have only a few weeks to build machine translation systems for language pairs that they have not previously worked on. We effectively demonstrated this with our shared task. For instance, seven teams built Finnish–English machine translation systems, a language pair that was certainly not of their immediate concern before.

In contrast to the bigger NIST competition, we wanted to keep the barrier of entry as low as possible. We provided not only training data from the Europarl corpus (Koehn, 2005), but also additional resources: sentence and word alignments, the decoder Pharaoh[1] (Koehn, 2004b), and a language model, so that participation was feasible even as a graduate level class project.

Using about 15 million words of translated text, participants were asked to build a phrase-based statistical machine translation system. The focus of the task was to build a probabilistic phrase translation table, since most of the other resources were provided — for more on phrase-based statistical machine translation, refer to Koehn et al. (2003). The participants' systems were compared by how well they translated 2000 previously unseen test sentences from the same domain.

The shared task operated within an extremely short timeframe. The workshop and hence the shared task was accepted on February 22, 2005 and announced on March 3. The official test data was made available on April 3, results were due one week later. Despite this tight schedule, eleven research groups participated and built a total of 32 machine translation systems for the four language pairs.

1 Goals

When setting up this competition, we were motivated by a number of goals. We set out to:

Create a platform to demonstrate the effectiveness of novel ideas: The research community is easily balkanized, where different groups work on

[1] http://www.isi.edu/licensed-sw/pharaoh/

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 119–124,
Ann Arbor, June 2005.

different data sets and under different conditions, so that it becomes often hard to assess, how effective a novel method is. By creating an environment with common test and training sets, language model, preprocessing, and even decoder, the effect of other model choices can be more easily demonstrated.

Work on new language pairs, new problems: Different language pairs pose different challenges. We picked Finnish–English and German–English for the special problems of rich morphology, word order, which are a challenge to current phrase-based SMT methods.

Enable more researchers to get engaged in SMT research: One of our main goals with providing as many resources as possible was to keep the barrier of entry low. Participants could use the word alignment and other resources and focus on phrase extraction. We hoped to attract researchers that are relatively new to the field. We were satisfied to learn that many entries are by graduate students working on their own.

Promote and create free resources: Academic research thrives on freely available resources. The field of statistical machine translation has been blessed with a long tradition of freely available software tools — such as GIZA++ (Och and Ney, 2003) — and parallel corpora — such as the Canadian Hansards[2]. Following this lead, we made word alignments and a language model available for this competition in addition to our previously published resources (Europarl and Pharaoh). The competition created resources as well. Most teams agreed to share system output and their model files. You can download them from the competition web site[3].

Promote work on European language pairs: Finally, we wanted to promote work on European languages. The increasing economic and political ties within the European Union create a huge need for translation services. We would like to see researchers rise to the challenge of creating high quality machine translation systems to fill these needs.

We are very grateful for the strong participation, especially by researchers who are relatively new to the field.

[2]http://www.isi.edu/natural-language/download/hansard/

[3]http://www.statmt.org/wpt05/mt-shared-task/

2 Rules of Engagement

We set up a machine translation competition for four language pairs. We chose Spanish–English and French–English, because many researchers would be familiar with these languages. We chose German–English for its special problems with word order (such as nested constructions and split verb groups) and morphology. Finally, we picked Finnish–English for the rich agglutinative morphology of Finnish.

Statistical machine translation systems are typically trained on sentence-aligned parallel corpora. We selected Europarl[4], a freely available parallel corpus in eleven languages. In addition, we also made a word alignment available, which was derived using a variant of the current default method for word alignment – Och and Ney (2003)'s refined method.

Figure 1 details some properties of the parallel corpora. The training corpus is most of the Europarl corpus, only the text of sessions from last quarter of the year 2000 was reserved for testing. The corpus has the size of roughly 15 million English words in 700,000 sentences – these numbers differ for each of the four parallel corpora due to the different number of discarded sentences during sentence alignment and after enforcing a 40 word length limit for sentences.

The number of foreign words differs even more dramatically. The effect of Finnish morphology manifests itself in a low number of words (just over 11 million), but a high number of distinct words (more than 5 times as many as in the English half).

The test corpus consists of 2000 sentences aligned across all five languages. Note that the output of each system is compared against the same English references for all source languages. The number of total words, distinct words, and words not seen in the training data reflects again the morphology effect.

For researchers willing to create their own word alignment, we suggested the use of GIZA++[5], an implementation of the IBM word-based machine translation models, which also assisted the creation of the provided word alignments.

We trained a language model on the English part

[4]http://www.statmt.org/europarl/

[5]http://www.fjoch.com/GIZA++.html

	Spanish–English	French–English	Finnish–English	German–English
	Training corpus			
Sentences	730,740	688,031	716,960	751,088
Source words	15,676,710	15,323,737	11,318,287	15,256,793
English words	15,222,105	13,808,104	15,492,903	16,052,269
Distinct source words	102,886	80,349	358,345	195,291
Distinct English words	64,123	61,627	64,662	65,889
	Test corpus			
Sentences	2,000			
Source words	60,276	65,029	41,431	54,247
English words	57,945			
Distinct source words	7,782	7,285	11,996	8,666
Distinct English words	6,054			
Unseen source words	209	143	737	377

Figure 1: Properties of the Europarl training and test corpora used in the shared task

of the Europarl corpus using the SRI language modeling toolkit (Stolke, 2002). Finally, we suggested the use of Pharaoh (Koehn, 2004b), a phrase-based machine translation decoder.

How well does this setup match the state of the art? The MIT system using the Pharaoh decoder (Koehn, 2004a) proved to be very competitive in last year's NIST evaluation. However, the field is moving fast, and a number of steps help to improve upon the provided baseline setup, e.g., larger language models trained on general text (up to a billion words have been used), better reodering models (e.g., suggested by Tillman (2004) and Och et al. (2004)), better language-specific preprocessing (Koehn and Knight, 2003) and restructuring (Collins et al., 2005), additional feature functions such as word class language models, and minimum error rate training (Och, 2003) to optimize parameters.

Some of these steps (e.g., improved reordering models) go beyond the current capabilities of Pharaoh. However, we are hopeful that freely available software continues to match or at least follow closely the state of the art.

We announced the shared task on March 3, and provided all the resources mentioned above (also a development test corpus to track the quality of systems being developed). The test schedule called for the translation of 2000 sentence for each of the four language pairs in the week between April 3–10. We allowed late submissions up to April 17.

3 Results

Eleven teams from eight institutions in Europe and North America participated, see Figure 2 for a complete list. The figure also indicates, if a team used the Pharaoh decoder (eight teams), the provided language model (seven teams) and the provided word alignment (four did, three of those with additional preprocessing or additional data).

Translation performance was measured using the BLEU score (Papineni et al., 2002), which measures n-gram overlap with a reference translation. In our case, we only used a single reference translation, since the test set was taken from a held-out portion of the Europarl corpus. On the other hand we used a relatively large number of test sentences to guarantee that the BLEU results are stable despite the fact that we used only one reference translation for each sentence.

Shared tasks like this one, of course, bring out the competitive spirit of participants and can draw criticisms about being a horse race. From an outside perspective, however, it is far more interesting to learn which methods and ideas proved to be successful, than who won the competition.

Taking stock of the results — see Figure 3 — one observes a very packed field at the top. While the participants from the University of Washington produced the best translations for every single language pair, the distance to many other participant scores

ID	Team	Pharaoh	LM	Word Al.
cmu-b	Carnegie Mellon University, USA - Bing Zhao	yes	yes	no
cmu-j	Carnegie Mellon University, USA - Ying (Joy) Zhang	yes	yes	no
glasgow	University of Glasgow, UK	yes	yes	yes+
nrc	National Research Council, Canada	no	no	no
rali	University of Montreal / RALI, Canada	yes	yes	no
saar	Saarland University, Germany	yes	yes	yes
uji	University Jaume I, Spain	yes	yes	yes+
upc-j	Polytechnic University of Catalonia, Spain - Jesus Gimenez	yes	yes	no
upc-m	Polytechnic University of Catalonia, Spain - Marta Ruiz	no	no	no
upc-r	Polytechnic University of Catalonia, Spain - Rafael Banchs	no	no	no
uw	University of Washington, USA	yes	no	yes+

Figure 2: The eleven participating teams: the table also lists, if the Pharaoh decoder, the provided language model, and the provided word alignment was used (yes+ indicates additional preprocessing)

is within a BLEU percentage point or two. As one might have expected, the scores are best for Spanish and French, and worst for Finnish. Figure 4 shows some typical output of the submitted systems.

The proceedings to the workshop include detailed system descriptions of all participants. Novel phrase extraction approaches were proposed, along with better preprocessing, language modeling, rescoring, and other ideas. We are certain that better performance can be achieved by combining some of the methods used by different participants.

And hence, we would like to pose the challenge to the research community to build and test better systems using the provided resources. We will gladly list additional results on the competition web site.

4 Survey

Following the end of the competition, we sent out a questionnaire to the participants. One of the questions what they would like to see different in a potential future competition.

We listed four potential changes: 70% of the respondends checked *translation from English*, 50% checked *out of domain test data*, 40% checked *more language pairs*, 0% checked *fewer language pairs*.

Additional suggestions were: alternatives to the BLEU scoring method (maybe human judgment by participants themselves), transitive translation using pivot languages, translation of resource-poor languages, and more time to prepare for the task.

5 Outlook

Given the short timeframe, one should view the system performances (albeit very competitive with the state of the art) as a baseline effort on the task of open domain text translation between European languages.

We hope that future researchers will use the provided environment as a test bed for their machine translation systems. We will continue to publish any scores reported to us.

Since we placed much of the systems' output online, the interested reader may be inspired to more closely explore the quality and shortcomings. Even some of the model files have been made available, so it is even possible to download and install some of the systems.

References

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL'05), Main Volume*, pages 531–540, Ann Arbor, Michigan.

Koehn, P. (2004a). The foundation for statistical machine translation at MIT. In *Proceedings of Machine Translation Evaluation Workshop 2004*.

Koehn, P. (2004b). Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation*

Spanish-English

System	BLEU	1/2/3/4-gram precision (bp)
uw	30.95	64.1/36.6/24.0/16.3 (1.000)
upc-r	30.07	63.1/35.8/23.2/15.6 (1.000)
upc-m	29.84	63.9/35.5/23.0/15.5 (0.995)
nrc	29.08	62.7/34.9/22.2/14.7 (1.000)
rali	28.49	62.4/34.5/21.9/14.4 (0.992)
upc-j	28.13	61.5/33.8/21.4/14.1 (1.000)
saar	26.69	61.0/33.1/20.7/13.5 (0.973)
cmu-j	26.14	61.2/32.4/19.8/12.6 (0.986)
uji	21.65	59.7/27.8/15.2/8.7 (1.000)

French-English

System	BLEU	1/2/3/4-gram precision (bp)
uw	30.27	64.8/36.8/23.8/16.0 (0.981)
upc-r	30.20	63.9/36.2/23.3/15.6 (0.998)
nrc	29.53	63.7/35.8/22.7/14.9 (0.997)
rali	28.89	62.6/34.7/22.0/14.6 (1.000)
cmu-b	27.65	63.1/34.0/20.9/13.3 (0.995)
cmu-j	26.71	61.9/33.0/20.3/13.1 (0.984)
saar	26.29	60.8/32.5/20.1/12.9 (0.982)
glasgow	23.01	57.3/28.0/16.7/10.5 (1.000)
uji	21.25	59.8/27.7/14.8/8.3 (1.000)

Finnish-English

System	BLEU	1/2/3/4-gram precision (bp)
uw	22.01	59.0/28.6/16.1/9.4 (0.979)
nrc	20.95	57.8/27.2/14.8/8.4 (0.996)
upc-r	20.31	56.6/26.0/14.3/8.3 (0.993)
rali	18.87	55.2/24.7/13.1/7.1 (0.998)
saar	16.76	58.4/26.3/14.2/8.0 (0.819)
uji	13.79	60.0/23.2/10.8/5.3 (0.821)
cmu-j	12.66	53.9/21.7/10.7/5.7 (0.775)

German-English

System	BLEU	1/2/3/4-gram precision (bp)
uw	24.77	62.2/31.8/18.8/11.7 (0.965)
upc-r	24.26	59.7/30.1/17.6/11.0 (1.000)
nrc	23.21	60.3/29.8/17.1/10.3 (0.979)
rali	22.91	58.9/29.0/16.8/10.3 (0.982)
saar	20.48	58.0/27.5/15.5/9.2 (0.938)
cmu-j	18.93	59.2/26.8/14.3/8.1 (0.914)
uji	18.89	59.3/25.5/13.0/7.2 (0.976)

Figure 3: The scores for the participating systems (BLEU and its components n-gram-precision and brevity penalty)

in the Americas, AMTA, Lecture Notes in Computer Science. Springer.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit X (submitted)*.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Stolke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

Tillman, C. (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Reference
We know all too well that the present Treaties are inadequate and that the Union will need a better and different structure in future , a more constitutional structure which clearly distinguishes the powers of the Member States and those of the Union .

Input Spanish
Sabemos muy bien que los Tratados actuales no bastan y que , en el futuro , será necesario desarrollar una estructura mejor y diferente para la Unión Europea , una estructura más constitucional que también deje bien claras cuáles son las competencias de los Estados miembros y cuáles pertenecen a la Unión .
Best system (Spanish–English)
we all know very well that the current treaties are not enough and that , in the future , it will be necessary to develop a structure better and different for the european union , a structure more constitutional also make it clear what the competences of the member states and what belongs to the union .
Worst System (Spanish–English)
we know very well that the current treaties not enough and that , in the future , will be necessary develop a better structure and different to the european union , a structure more constitutional that also be well clear the powers of the member states and what belong to the union .

Input French
Nous savons très bien que les Traités actuels ne suffisent pas et qu ' il sera nécessaire à l ' avenir de développer une structure plus efficace et différente pour l ' Union , une structure plus constitutionnelle qui indique clairement quelles sont les compétences des états membres et quelles sont les compétences de l ' Union .
Best system (French–English)
we know very well that the current treaties are not enough and that it will be needed in the future to develop a structure more effective and different for the union , a structure more constitutional which clearly indicates what are the competence of member states and what are the powers of the union .

Input Finnish
Tiedämme oikein hyvin , että nykyiset perustamissopimukset eivät ole riittäviä ja että tulevaisuudessa on tarpeen kehittää unionille parempi ja toisenlainen rakenne , siis perustuslaillisempi rakenne , jossa mys ilmaistaan selkeämmin , mitä jäsenvaltioiden ja unionin toimivaltaan kuuluu
Best system (Finnish–English)
we know very well that the existing founding treaties do not need to be developed for the union and a different structure , therefore perustuslaillisempi structure , which also expresses clearly what the member states and the union 's competence is not sufficient and that better in the future .

Input German
Uns ist sehr wohl bewusst , dass die geltenden Verträge unzulänglich sind und künftig eine andere , effizientere Struktur für die Union entwickelt werden muss , nämlich eine stärker konstitutionell ausgeprägte Struktur mit einer klaren Abgrenzung zwischen den Befugnissen der Mitgliedstaaten und den Kompetenzen der Union .
Best system (German–English)
the union must be developed , with a major institutional structure with a clear demarcation between the powers of the member states and the competences of the union is well aware that the existing treaties are inadequate and in the future , a different , more efficient structure for us .

Figure 4: The first sentence of the test corpus and system translations