Comparative Evaluation of Research vs. Online MT Systems

Antonio Toral[†], Federico Gaspari[†], Sudip Kumar Naskar[†*] and Andy Way[†*]
CoSyne[†] / CNGL[*]
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland
{atoral,fgaspari,snaskar,away}@computing.dcu.ie

Abstract

This paper reports MT evaluation experiments that were conducted at the end of year 1 of the EU-funded CoSyne[1] project for three language combinations, considering translations from German, Italian and Dutch into English. We present a comparative evaluation of the MT software developed within the project against four of the leading free web-based MT systems across a range of state-of-the-art automatic evaluation metrics. The data sets from the news domain that were created and used for training purposes and also for this evaluation exercise, which are available to the research community, are also described. The evaluation results for the news domain are very encouraging: the CoSyne MT software consistently beats the rule-based MT systems, and for translations from Italian and Dutch into English in particular the scores given by some of the standard automatic evaluation metrics are not too distant from those obtained by well-established statistical online MT systems.

1 Introduction

CoSyne is an EU-funded project that aims at facilitating the synchronization across different languages of the contents of wiki sites. This is a particularly challenging endeavour, because of the conflation of dynamic user-generated (or user-edited/corrected) content and multilingual aspects. Today, wikis are regarded as very popular and efficient tools by the public and Internet users as a whole, as well as in specific scenarios such as large organizations, companies and also smaller groups of people who want to access shared, collaboratively built and open-ended repositories of collective knowledge and information.

Currently, wikis that offer information in multiple languages or that serve international or multilingual communities, rely on users themselves to manually translate wiki entries on the same subject. This is not only a time-consuming process, but also the source of many inconsistencies, as users update the different language versions separately, and every update is likely to increase the divergence in terms of content of the multilingual entries. This results in an obvious loss of information for all but the most widely used languages, which tend to have much more information than the others. Conversely, information that is available in less popular languages is unlikely to be translated into other versions of the wiki site, and therefore will remain confined to the smaller user communities, without being accessible to the wider population of wiki users.

The overall aim of the CoSyne project is to automate the dynamic multilingual content synchronization process of wiki sites across languages, by achieving robust MT of noisy user-generated content between 6 languages (consisting of 4 core languages and 2 languages with limited resources to demonstrate adaptability of the system). The three language pairs covered in year 1 of the project (March 2010-February 2011) are Dutch-English, German-English, and Italian-English, and we report on them in this paper. Later in the project, Turkish and Bulgarian will also be added, to show the adaptability of the system.

The CoSyne system will be integrated via web services with the open-source MediaWiki[2] package, which is the most commonly used wiki plat-

[1] http://www.cosyne.eu/

[2] http://www.mediawiki.org/

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 13–20
Leuven, Belgium, May 2011

form. The overall CoSyne system will include several components to help deal with typically noisy and largely unpredictable wiki content (e.g. document structure modeling and document structure induction, a textual entailment module, etc.), but this paper focuses exclusively on the initial evaluation of its MT software, as one of its key components, against state-of-the-art systems used as baselines.

The CoSyne consortium consists of 7 partners from 4 different EU countries: Germany, Ireland, Italy and the Netherlands. The consortium includes 3 academic partners: University of Amsterdam, Fondazione Bruno Kessler and Dublin City University (DCU); 1 research organization: Heidelberg Institute for Theoretical Studies, and 3 end-users: Netherlands Institute for Sound and Vision (NISV), Deutsche Welle (DW) and Vereniging WikiMedia Nederland. The academic and research partners ensure the cutting-edge interest of the project, emphasizing the need to go beyond the state-of-the-art in order to tackle the problems addressed in this project. At the same time, the participation of end-users ensures that all research outcomes of the project are relevant to the wider society and boosts the potential for the uptake of the results beyond academia.

In this paper we focus on an initial comparative evaluation of the CoSyne MT software (a statistical system developed by the University of Amsterdam, the project coordinator), against four of the leading free web-based MT systems for three language combinations over three data sets from the news domain across a range of state-of-the-art automatic evaluation metrics. This is part of the ongoing effort to evaluate the effectiveness and quality of the MT component developed within the project, assess its capabilities compared to widely used MT software, and monitor the progress of the system throughout the duration of the project, measuring its improvement with regular evaluation cycles.

The rest of the paper is structured as follows. Section 2 introduces the metrics that have been used to evaluate the performance of the MT systems in comparative terms. The data used for the evaluation along with the procedure used to select it are briefly presented in Section 3. Section 4 gives a description of the MT software under evaluation and of the other systems used as baselines for the comparative evaluation. Evaluation results are presented and discussed in Section 5. Finally, in Section 6 we draw some conclusions and outline plans for future work as part of the CoSyne project.

2 Metrics

This section presents an up-to-date overview of some of the most widely used automatic MT evaluation metrics, discussing their advantages as well as drawbacks, laying particular emphasis on the metrics used in the comparative evaluation presented in Section 5. The performance of the CoSyne MT system in the early stages of its development can be measured, and its improvement can be monitored over time, against these standard metrics in a reliable and replicable fashion.

To ensure the best possible coverage, we decided to use a wide array of metrics, particularly those judged best by recent meta-evaluation exercises (e.g. Callison-Burch et al., 2007; Callison-Burch et al., 2010), without confining ourselves to prominent n-gram based metrics. Since there is no consensus on a single individual metric which is thought to accurately measure MT performance, we decided to adopt an inclusive approach, considering the results of a variety of measures. This should provide a picture that is as reliable and fine-grained as possible.

One of the most widely used automatic MT evaluation metrics is BLEU (Papineni et al., 2002), a string-based metric which has come to represent something of a de facto standard in the last few years. This is not surprising given that today most MT research and development efforts are concentrated on statistical approaches; BLEU's critics argue that it tends to favour statistical systems over rule-based ones (Callison-Burch et al., 2006). Using BLEU is fast and intuitive, but while this metric has been shown to produce good correlations with human judgment at the document level (Papineni et al., 2002), especially when a large number of reference translations are available, correlation at sentence level is generally low.

The NIST evaluation metric (Doddington, 2002) is also string-based, and gives more weight in the evaluation to less frequent n-grams. While this metric has a strong bias in favour of statistical systems, it provides better adequacy correlation than BLEU (Callison-Burch et al., 2006).

The GTM metric (Turian et al., 2003) is based on standard measures adopted in other NLP applications (precision, recall and F-measure), which makes its use rather straightforward for NLP practitioners. It focuses on unigrams and rewards sequences of correct unigrams, applying moderate penalties for incorrect word order.

METEOR (Banerjee and Lavie, 2005) uses stemming and synonymy relations to provide a

more fine-grained evaluation at the lexical level, which reduces its bias towards statistical systems. One drawback of this metric is that it is language-dependent since it requires a stemmer and WordNet,[3] and it can currently be applied in full only to English, and partly to French, Spanish and Czech, due to the limited availability of synonymy and paraphrase modules. METEOR-NEXT (Denkowski and Lavie, 2010) is an updated version of the same metric.

The TER metric (Snover et al., 2006) adopts a different approach, in that it computes the number of substitutions, insertions, deletions and shifts that are required to modify the output translation so that it completely matches the reference translation(s). Its results are affected less by the number of reference translations than is the case for BLEU, and the rationale behind this evaluation metric is quite simple to understand for people who are not MT experts, as it provides an estimation of the amount of post-editing effort needed by an end-user. Another metric based on error rates which preceded TER is WER (Nießen et al., 2000). We omitted WER and its extension mWER (Nießen et al., 2000) from the experiments reported here as they seem to have been superceded by more recent metrics.

TER-plus (Snover et al., 2009) is an extension of TER using phrasal substitutions relying on automatically generated paraphrases, stemming, synonyms and relaxed shifting constraints. This metric is language-dependent and requires WordNet. It has been shown to have the highest average rank in terms of Pearson and Spearman correlation (Przybocki et al., 2008).

The DCU-LFG metric (Owczarzak et al., 2007) exploits LFG dependencies and has only a moderate bias towards statistical systems. It requires a dependency parser.

It should be noted that among the above measures, METEOR, METEOR-NEXT, TER-plus and DCU-LFG can only be used for English as a target language at the present time, given the language-specific resources that they require.

3 Data

Sections 3.1—3.3 describe the data created and used for training as well as evaluation purposes for each language pair (German—English, Italian—English and Dutch—English, respectively) and the procedures followed to derive test sets from that data. For each of these language pairs we aligned approximately 2,000 sentence pairs broadly coming from the news domain; 1,000 were used for training purposes, while the remaining 1,000 sentence pairs were used for the evaluation presented here. The test sets were designed and built in accordance with the requirements put forward by the end-user consortium partners - DW, NISV and the Wikimedia Foundation Netherlands, so as to match as closely as possible their envisaged use of the CoSyne system, and in particular of its MT component in their wiki settings. In particular, DW and NISV rely heavily on wiki sites to circulate and share information both within their organizations as well as externally to the audiences and publics that they serve in multiple languages.

3.1 German—English

The initial input for this language pair was provided by DW. It consisted of documents in RDF[4] format coming from two online journals: Europa Aktuell 2001 to 2010 (2,201 documents) and Global 3,000 (80 documents).

DW provided this particular dataset as raw input for testing and evaluation. These records were adapted to turn them in the appropriate XML format required for further automated processing. The items were originally not necessarily in the same order of appearance in both language versions; as this was a requirement, the XML files were run through a script to identify the differences, subsequently adapted using the Stylus Studio XML adaptation program, and amended to finally obtain parallel language versions (with corresponding items in the same order). The final version was delivered by DW in the XML format as specified by DCU who took care of aligning the data and using them to run the actual evaluations.

A Perl script was developed to extract titles and running text from pairs of parallel documents. Apart from extracting the contents the script also invokes:

- TreeTagger,[5] to sentence split and lemmatise the documents.
- Hunalign,[6] with a bilingual dictionary derived from Apertium's English—

[3] http://wordnet.princeton.edu/

[4] http://www.w3.org/RDF/

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http://mokk.bme.hu/resources/hunalign

German dictionary[7] to sentence align the documents using the lemmatized versions produced by TreeTagger.

Another Perl script was developed to enrich Hunalign output, consisting at this step of aligned sentence pairs, with several additional pieces of information:

- alignment score at sentence level
- alignment score at document level
- number of words in the source sentence
- number of words in the target sentence
- whether the sentence consists of a news item title or running text
- identifier of the document pair

Based on these factors we chose a test corpus with sentences that fulfilled these criteria:

- The score of the sentence alignment is above the threshold 0.7. This is to ensure that the translations are accurate.
- The score of the global alignment (at document level) is above the threshold 0.7. This ensures that the sentence pairs are extracted from highly parallel document pairs.
- The difference in length (number of words) between the source and target sentence expressed as a percentage is below 25%. This is to ensure that the versions in the two languages match quite closely.
- The minimum number of words in the source and target sentences is 4. This prevents the test set from containing very short sentences.
- The percentage of sentences that consist of titles and sentences that are part of running text are preserved in the test set. We allow a maximum 2% variation with respect to the distribution in the whole data set, where 19.35% of the sentences are titles.

While the values for these parameters have not been determined empirically to be necessarily the best overall settings, manual investigation of the test sets confirms these values to be effective in practice. With these values, only 1,903 sentence pairs are extracted. Therefore we decided to slightly decrease the value of the global alignment quality from 0.7 to 0.68, as by doing so we can extract the planned 2,000 sentence pairs.

The resulting test set contains 2,000 aligned sentences, including titles of the articles (these eventually account for 21.35% of the pairs). The English file has 26,938 words and average sentence length is 13.47 words. The German file has slightly less content, with 25,797 words, and average sentence length is 12.9 words.

3.2 Italian—English

The source for the English—Italian test set was the AsiaNews website,[8] which provides news on current events in Italian and English. The news texts included in the test set have been posted online in the last few years, up to July 2010. 87 document pairs were extracted manually, and the sentences were stored in two separate plain-text files: sentence number X in the Italian file is the translation of sentence number X in the English file.

The titles of the news article were included along with the running text of the news report, while all other elements were excluded: picture captions, names of the authors, indications of place names at the beginning of the report, etc.

Sentence splitting was done manually, with a new sentence (new line) created every time there was a full stop at the end of a sentence (i.e. no split was inserted when the full stop was found in abbreviations like “Mr. Smith” or “Gen. Ross”). Wherever possible the same punctuation marks were kept in both versions, harmonizing them manually (e.g. a full stop was inserted instead of a semicolon in English, where the Italian sentence ended with a full stop). Intra-sentence commas were not modified, leaving them as they were. Translations found on the website were usually very good, but varied in that some texts were more faithful (with a closely matching structure in terms of paragraphs, sentence organization, etc.), while in other cases the two versions were occasionally rather different (omitted sentences, shorter paragraphs with only summarized information, etc.). An effort was made to include in the aligned data set only sentences that are highly symmetrical.

Where the sentences diverged, both of them were omitted, or one of them was amended as necessary to make the pair of aligned segments more “similar” in form and structure. We estimate that around 20% of the content in the original bilingual versions of the articles was elimi-

[7] http://apertium.svn.sourceforge.net/viewvc/apertium/incubator/apertium-de-en/

[8] http://www.asianews.it

nated or slightly amended for the sake of having highly parallel sentences.

The resulting test set contains 1,965 aligned sentences, including titles of the articles (these account for 4.4% of the sentence pairs). The Italian file has 38,607 words (on average 444 words per document) and average sentence length is 19.3 words. The English file has slightly less content, 38,090 words, on average 438 words per document, and average sentence length is 19 words.

3.3 Dutch—English

NISV provided three different data sets:

- België Diplomatie consists of 418 HTML document pairs extracted from the Belgian Foreign Affairs website.[9]
- Video Active is an XML file containing 1,076 document pairs concerning the description of television programs.[10]
- NISV wiki has 30 document pairs consisting of pages from the NISV wiki.[11]

NISV extracted the HTML pages from the België Diplomatie website with a Python script. Being a partner in the Video Active project, NISV had access to its metadata repository and created a Java program to extract this metadata and store it in XML files. Apart from providing the translated wiki pages, NISV also manually created easy-to-process Dutch—English sentence pairs in separate (plain text) files.

Three Perl scripts were developed to extract titles and running text from pairs of parallel documents of the België Diplomatie, Video Active and NISV wiki data sets. The strategy followed is along the lines of that previously described for German—English (Section 3.1). The bilingual dictionary used comes from Apertium's Dutch—English dictionary.[12]

With the values introduced (Section 3.1.), only 1,902 sentence pairs are extracted. Therefore we decided to slightly decrease the value of the global alignment quality from 0.7 to 0.66, as by doing so we can extract the planned 2,000 sentence pairs. The derived set contains 350 sentence pairs (17.5%) from the NISV wiki data, 618 (30.9%) from Video Active while the remaining 1,032 pairs (51.6%) come from the België Diplomatie data.

The resulting test set contains 2,000 aligned sentences, including titles of the articles (these account for 7.7% of the sentence pairs). The Dutch file has 45,546 words and average sentence length is 22.8 words. The English file has slightly more content, 46,390 words, and the average sentence length is 23.2 words.

4 MT Systems

The following four free online MT systems were used for the baseline evaluation of the CoSyne MT system developed by the University of Amsterdam (Martzoukos and Monz, 2010):

- Google Translate[13]
- Bing Translator[14]
- Systran[15]
- FreeTranslation[16]

These four online MT services were selected first of all because they all cover the three language pairs addressed in year 1 of the CoSyne project (German—English, Italian—English and Dutch—English in both directions). In addition, these are among the most popular free web-based MT systems and are heavily used by the general public of Internet users (Gaspari, 2006; Gaspari and Hutchins, 2007). A final consideration was that three of these five systems are statistical (CoSyne, Google Translate and Bing Translator), while the other two are rule-based (FreeTranslation and Systran). As a result, this mixture of systems offers a good picture of the MT quality currently offered by state-of-the-art representatives of both approaches.

5 Results

In what follows, the results of the MT evaluation carried out at the end of the first year of the CoSyne project (February 2011) are presented and discussed in this order: German—English (Section 5.1), Italian—English (5.2) and Dutch—English (5.3). For each of these 3 language pairs, data are included that show the comparison between the CoSyne MT software and state-of-the-art MT systems serving as benchmarks of statistical (Google and Bing) as well as rule-based (Systran and FreeTranslation) approaches across a range of well-established automatic MT

[9] http://diplomatie.belgium.be
[10] http://www.videoactive.eu
[11] http://www.beeldengeluidwiki.nl
[12] http://apertium.svn.sourceforge.net/viewvc/apertium/incubator/apertium-en-nl/
[13] http://translate.google.com
[14] http://www.microsofttranslator.com
[15] http://www.systran.co.uk/
[16] http://www.freetranslation.com

evaluation metrics (discussed in Section 2). The data are presented in tables giving the actual numerical results of the evaluation, accompanied by figures to facilitate comparison, followed by a brief analysis of the most interesting findings. Finally, Section 5.4 provides a summary discussion of the whole evaluation experiment.

Results of statistical significance tests are also included to indicate the validity of the comparisons between the CoSyne MT software and the benchmark MT systems. Statistical significance is represented in the tables with characters written as superscripts. For each system and metric,[17] a character *n* means that the current score is significantly better than the system in column *n*, e.g., *c*,*d* indicates that the current score is better than those obtained by the systems in the third and fourth columns. Finally, it should be noted that to facilitate the interpretation of the results, the NIST score was divided by a factor of 10 for the sake of consistency in the presentation. Similarly, TERp and TER scores are indicated as *1-x* to reverse the trend and make it more comparable to the other metrics.

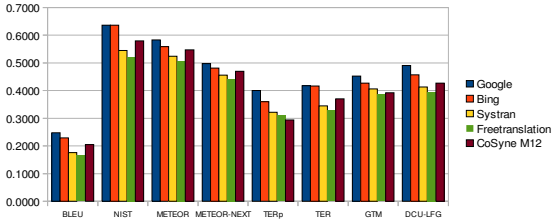
Regarding the software used to carry out the evaluation, we used the following implementations:

– BLEU and NIST: mteval11b-sig.pl,
– METEOR: meteor 1.0,
– METEOR-NEXT: meteor 1.2,
– TER: terp 0.1 (binary terp_ter),
– TERp: terp 0.1 (binary terpa),
– GTM: gtm 1.4,
– DCU-LFG: version submitted to Metrics MATR 2010 (He et al., 2010),
– Statistical significance tests: ARK's code[18] (BLEU and NIST) and FastMtEval[19] (GTM). P-value is set to 0.01.

5.1 German into English

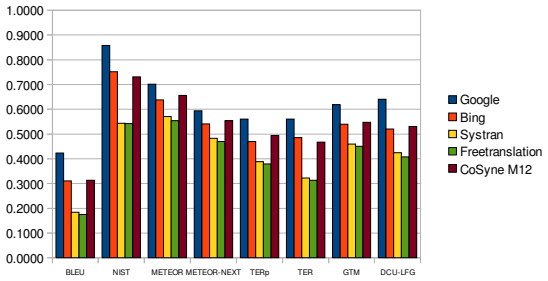
For German—English translation, in most cases (with the exceptions of the TERp and GTM metrics) the quality of the CoSyne MT software is between the statistical MT systems (Google and Bing), which show a better performance, and the rule-based ones (Systran and FreeTranslation), which tend to be outperformed by the CoSyne MT software. Systran outperforms FreeTranslation across all the metrics, and in most cases Google receives better evaluations than Bing (except for NIST, where Bing has a very tiny lead, and the two TER scores are very similar). TERp ranks the CoSyne MT system fifth, slightly below Systran and FreeTranslation, while the GTM score of the CoSyne MT system is slightly better than FreeTranslation, but not as good as Systran. Although the global picture is somewhat fragmented, in general the month 12 implementation of the CoSyne MT software performs better than the rule-based systems, but not yet as well as the statistical benchmark systems used as baselines. It is worth mentioning that these systems have undergone several years of extensive development, relying on massive amounts of resources.

de-en	*Google*	*Bing*	*Systran*	*Freetranslation*	*CoSyne M12*
BLEU	0.2477^{b,c,d,e}	0.2294^{c,d}	0.1752^{d}	0.1657	0.2052^{c,d}
NIST	0.6358^{b,c,d,e}	0.6362^{c,d,e}	0.5447^{d}	0.5212	0.5788^{c,d}
METEOR	0.5830	0.5584	0.5239	0.5060	0.5470
METEOR-NEXT	0.4977	0.4807	0.4552	0.4422	0.4692
TERp	0.4000	0.3600	0.3216	0.3100	0.2941
TER	0.4172	0.4161	0.3444	0.3273	0.3700
GTM	0.4517^{b,c,d,e}	0.4270^{c,d,e}	0.4057^{d,e}	0.3849	0.3914
DCU-LFG	0.4899	0.4570	0.4133	0.3957	0.4261



5.2 Italian into English

it-en	*Google*	*Bing*	*Systran*	*Freetranslation*	*CoSyne M12*
BLEU	0.4235^{b,c,d,e}	0.3106^{c,d}	0.1840^{d}	0.1754	0.3137^{c,d}
NIST	0.8579^{b,c,d,e}	0.7517^{c,d,e}	0.5439^{d}	0.5427	0.7318^{c,d}
METEOR	0.7017	0.6384	0.5709	0.5537	0.6565
METEOR-NEXT	0.5942	0.5412	0.4832	0.4700	0.5545
TERp	0.5600	0.4700	0.3890	0.3800	0.4946
TER	0.5599	0.4857	0.3225	0.3128	0.4679
GTM	0.6187^{b,c,d,e}	0.5394^{c,d}	0.4596^{d}	0.4510	0.5475^{c,d}
DCU-LFG	0.6400	0.5200	0.4244	0.4080	0.5311



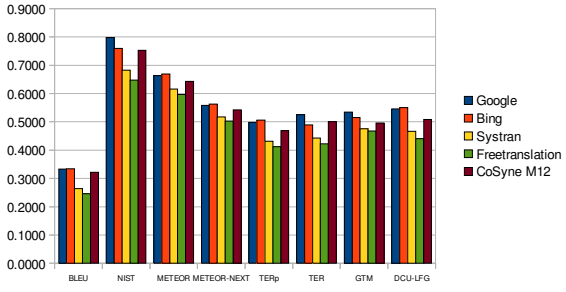
For Italian—English translation, Google consistently has the best performance across all the automatic evaluation metrics. Interestingly, the CoSyne MT system and Bing show similar performance, with the CoSyne software giving better results than Bing for some metrics (BLEU, METEOR, METEOR-NEXT, TERp, GTM and

[17] Statistical significance tests have been carried out for these metrics: BLEU, NIST and GTM.
[18] http://www.ark.cs.cmu.edu/MT/
[19] http://www.computing.dcu.ie/~nstroppa/index.php?page=softwares

DCU-LFG), which is a particularly encouraging result. The two rule-based systems (Systran and FreeTranslation) receive very similar scores for all evaluation metrics, showing much poorer performance than the statistical MT software (including the CoSyne MT system).

5.3 Dutch into English

nl-en	Google	Bing	Systran	Freetranslation	CoSyne M12
BLEU	0.3330[c,d,e]	0.3347[c,d,e]	0.2643[d]	0.2456	0.3223[c,d]
NIST	0.7986[b,c,d,e]	0.7596[c,d]	0.6830[d]	0.6479	0.7532[c,d]
METEOR	0.6633	0.6695	0.6161	0.5964	0.6431
METEOR-NEXT	0.5583	0.5628	0.5180	0.5032	0.5419
TERp	0.4987	0.5066	0.4315	0.4123	0.4690
TER	0.5251	0.4892	0.4424	0.4221	0.5000
GTM	0.5339[b,c,d,e]	0.5156[c,d,e]	0.4761[d]	0.4672	0.4956[c,d]
DCU-LFG	0.5459	0.5507	0.4661	0.4411	0.5080



For the Dutch—English translation task, the three statistical MT systems consistently and clearly outperform Systran and FreeTranslation based on all the automatic evaluation metrics. Google outperforms Bing for only three of the metrics (NIST, TER and GTM), whereas for the others Bing receives the higher score. Interestingly, based on TER, the CoSyne MT system does better than Bing, but not as well as Google. Finally, for all metrics, the score obtained by the CoSyne MT system is much higher than those of both Systran and FreeTranslation, and not particularly distant from those achieved by the other two statistical MT systems, which positively reflects the quality that the software has achieved after the initial 12 months of development.

5.4 Discussion

Overall Google Translate receives the best scores consistently across the various metrics for all language pairs. Bing Translator and the CoSyne MT system perform similarly: their results are noticeably inferior to Google Translate's (Bing obtains the best score for some metrics, mainly in Dutch-to-English translation, but none of them are significantly better than for Google Translate), but significantly higher[20] than those offered by Systran and FreeTranslation. Within the various language directions, the performance of Systran is better than that of FreeTranslation according to all evaluation metrics for all the three language pairs. Finally, for all evaluation metrics in each of the language pairs, the three statistical systems (Google Translate, Bing Translator and CoSyne) receive much higher scores than the two rule-based systems: Systran and FreeTranslation.

[20] When interpreting the results, one should bear in mind the limitations of the metrics used, e.g. that BLEU and NIST have a high bias towards statistical MT.

6 Conclusions and Future Work

The results presented in Section 5 show the baseline evaluation in all the three language directions covered in year 1 of the CoSyne project across a range of widely used automatic MT evaluation metrics. This will allow the project members to measure the performance of the MT component of the CoSyne system against state-of-the-art MT software, and monitor progress over time. In particular, this baseline evaluation will make it possible to prioritize and focus efforts on the development and fine-tuning of language pairs and/or language directions needing improvement. By repeating evaluations based on the well-established metrics presented in Section 2 at regular intervals, the improvement of the CoSyne MT system will be gradually monitored and its overall success measured.

This evaluation study has shown that rule-based MT systems are outperformed by statistical MT systems for data from the news domain. Plans currently underway to extend the evaluation of the CoSyne MT system include the development of a methodology for diagnostic MT evaluation based on linguistic checkpoints, similar to the one presented in Zhou et al. (2008), who used an ad-hoc tool called Woodpecker.

Acknowledgements

This work has been funded in part by the European Commission through the CoSyne project FP7-ICT-4-248531.We are also grateful to Christof Monz for giving us access to the CoSyne MT system.

References

Banerjee, S. and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association*

of Computational Linguistics. Ann Arbor, Michigan, USA. 65-72.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. Proceedings of EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy. 249-256.

Callison Burch, C., C.S. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-)Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic. 136-158.

Callison-Burch , C., P. Koehn, C. Monz, C., K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and METRICS MATR*. Uppsala, Sweden. 17-53.

Denkowski, M. and A. Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. *Proceedings of the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*. Uppsala, Sweden. 339-342.

Doddington, G. 2002. Automatic evaluation of MT quality using n-gram co-occurrence statistics. *Proceedings of Human Language Technology Conference 2002*. San Diego, CA, USA. 138-145.

Gaspari, F. 2006. The Added Value of Free Online MT Services: Confidence Boosters for Linguistically-challenged Internet Users, a Case Study for the Language Pair Italian-English. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas "Visions for the Future of Machine Translation"*. Cambridge, Boston, Massachusetts, USA. 46-55.

Gaspari, F. and J. Hutchins. 2007. Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects. *Proceedings of the Machine Translation Summit XI*. Copenhagen, Denmark. 199-206.

He, Y., J. Du, A. Way, and J. van Genabith. 2010. The DCU dependency-based metric in WMT-MetricsMATR 2010. *Proceedings of the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*. Uppsala, Sweden. 349-353.

Lin, C.-Y. and F.J. Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain. 605-612.

Martzoukos, S. and C. Monz. 2010. The UvA system description for IWSLT 2010. *Proceedings of the 7th International Workshop on Spoken Language Translation*. Paris, France. 205-208.

Nießen, S., F.J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for Machine Translation: Fast evaluation for MT research. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece. 39-45.

NIST Multimodal Information Group. National Institute of Standards and Technology, USA. Available at http://www.itl.nist.gov/iad/mig//tests/metricsmatr/2008/doc/mm08_evalplan_v1.1.pdf (last accessed April 29, 2011).

Owczarzak, K., J. van Genabith, and A. Way. 2007. Dependency-based automatic evaluation for Machine Translation. *Proceedings of the HLT-NAACL 2007 Workshop on Syntax and Structure in Statistical Machine Translation*. Rochester, NY, USA. 86-93.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of Machine Translation. In *40th Annual Meeting of the Association of Computational Linguistics*. Philadelphia, PA, USA. 311-318.

Przybocki, M., K. Peterson, and S. Bronsart. 2008. The NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR): Evaluation Specification Document".

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the Conference of the Association for Machine Translation in the Americas Conference*. 223-231.

Snover, M., N. Madnani, B. Dorr, and R. Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics*. Athens, Greece. 259-268.

Turian, J.P., L. Shen, and I.D. Melamed. 2003. Evaluation of Machine Translation and its evaluation. *Proceedings of MT Summit 2003*. New Orleans, LA, USA. 386-393.

Zhou, M., B. Wang, S. Liu, M. Li, D. Zhang, and T. Zhao. 2008. Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points. *Proceedings of COLING 2008*. Manchester, UK, 18-22 August 2008. 1121-1128.