SemEval-2014 Task 5: L2 Writing Assistant

Maarten van Gompel, Iris Hendrickx, Antal van den Bosch
Centre for Language Studies,
Radboud University Nijmegen,
The Netherlands
proycon@anaproy.nl,
i.hendrickx@let.ru.nl,
a.vandenbosch@let.ru.nl

Els Lefever and Véronique Hoste
LT3,
Language and Translation Technology Team,
Ghent University,
Belgium
els.lefever@ugent.be,
veronique.hoste@ugent.be

Abstract

We present a new cross-lingual task for SemEval concerning the translation of L1 fragments in an L2 context. The task is at the boundary of Cross-Lingual Word Sense Disambiguation and Machine Translation. It finds its application in the field of computer-assisted translation, particularly in the context of second language learning. Translating L1 fragments in an L2 context allows language learners when writing in a target language (L2) to fall back to their native language (L1) whenever they are uncertain of the right word or phrase.

1 Introduction

We present a new cross-lingual and application-oriented task for SemEval that is situated in the area where Word Sense Disambiguation and Machine Translation meet. Finding the proper translation of a word or phrase in a given context is much like the problem of disambiguating between multiple senses.

In this task participants are asked to build a translation/writing assistance system that translates specifically marked L1 fragments in an L2 context to their proper L2 translation. This type of translation can be applied in writing assistance systems for language learners in which users write in a target language, but are allowed to occasionally back off to their native L1 when they are uncertain of the proper lexical or grammatical form in L2. The task concerns the NLP back-end rather than any user interface.

Full-on machine translation typically concerns the translation of complete sentences or texts from L1 to L2. This task, in contrast, focuses on smaller fragments, side-tracking the problem of full word reordering.

We focus on the following language combinations of L1 and L2 pairs: *English-German*, *English-Spanish*, *French-English* and *Dutch-English*. Task participants could participate for all language pairs or any subset thereof.

2 Task Description

We frame the task in the context of second language learning, yielding a specific practical application.

Participants build a *translation assistance system* rather than a full machine translation system. The L1 expression, a word or phrase, is translated by the system to L2, given the L2 context already present, including right-side context if available. The aim here, as in all translation, is to carry the semantics of the L1 fragment over to L2 and find the most suitable L2 expression given the already present L2 context.

Other than a limit on length (6 words), we do not pose explicit constraints on the kinds of L1 fragments allowed. The number of L1 fragments is limited to one fragment per sentence.

The task addresses both a core problem of WSD, with cross-lingual context, and a sub-problem of Phrase-based Statistical Machine Translation; that of finding the most suitable translation of a word or phrase. In MT this would be modelled by the translation model. In our task the full complexity of full-sentential translation is bypassed, putting the emphasis on the semantic aspect of translation. Our task has specific practical applications and a specific intended audience, namely intermediate and advanced second language learners, whom one generally wants to encourage to use their target language as much as possible, but who may often feel the need to fall back to their native language.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 36–44,
Dublin, Ireland, August 23-24 2014.

Currently, language learners are forced to fall back to a bilingual dictionary when in doubt. Such dictionaries do not take the L2 context into account and are generally more constrained to single words or short expressions. The proposed application would allow more flexible context-dependent lookups as writing progresses. The task tests how effectively participating systems accomplish this.

The following examples illustrate the task for the four language pairs we offer:

- Input (L1=English,L2=Spanish): *"Todo ello, **in accordance** con los principios que siempre hemos apoyado."*
  Desired output: *"Todo ello, **de conformidad** con los principios que siempre hemos apoyado."*

- Input (L1-English, L2=German): *"Das, was wir heute machen, **is essentially** ein Ärgernis."*
  Desired output: *"Das, was wir heute machen, **ist im Grunde genommen** ein Ärgernis."*

- Input (L1=French,L2=English): *"I **rentre à la maison** because I am tired."*
  Desired output: *"I **return home** because I am tired."*

- Input (L1=Dutch, L2=English): *"Workers are facing a massive **aanval op** their employment and social rights."*
  Desired output: *"Workers are facing a massive **attack on** their employment and social rights."*

The task can be related to two tasks that were offered in previous years of SemEval: Lexical Substitution (Mihalcea et al., 2010) and most notably Cross-lingual Word Sense Disambiguation (Lefever and Hoste, 2013).

When comparing our task to the Cross-Lingual Word-Sense Disambiguation task, one notable difference is the fact that our task concerns not just words, but also phrases. Another essential difference is the nature of the context; our context is in L2 instead of L1. Unlike the Cross-Lingual Word Sense Disambiguation task, we do not constrain the L1 words or phrases that may be used for translation, except for a maximum length which we set to 6 tokens, whereas Lefever and Hoste (2013) only tested a select number of nouns. Our task emphasizes a correct meaning-preserving choice of words in which translations have to fit in the L2 context. There is thus a clear morphosyntactic aspect to the task, although less prominent than in full machine translation, as the remainder of the sentence, already in L2, does not need to be changed. In the Cross-Lingual Word Sense Disambiguation tasks, the translations/senses were lemmatised. We deliberately chose a different path that allows for the envisioned application to function directly as a translation assistance system.

A pilot study was conducted to test the feasibility of the proposed translation system (van Gompel and van den Bosch, 2014). It shows that L2 context information can be a useful cue in translation of L1 fragments to L2, improving over a non-context-informed baseline.

3 Data

We did not provide training data for this task, as we did not want to bias participating systems by favouring a particular sort of material and methodology. Moreover, it would be a prohibitively large task to manually collect enough training data of the task itself. Participants were therefore free to use any suitable training material such as parallel corpora, wordnets, or bilingual lexica.

Trial and test data has been collected for the task, both delivered in a simple XML format that explicitly marks the fragments. System output of participants adheres to the same format. The trial set, released early on in the task, was used by participants to develop and tune their systems on. The test set corresponds to the final data released for the evaluation period; the final evaluation was conducted on this data.

The trial data was constructed in an automated fashion in the way described in our pilot study (van Gompel and van den Bosch, 2014). First a phrase-translation table is constructed from a parallel corpus. We used the Europarl parallel corpus (Koehn, 2005) and the Moses tools (Koehn et al., 2007), which in turn makes use of GIZA++ (Och and Ney, 2000). Only strong phrase pairs (exceeding a set threshold) were retained and weaker ones were pruned. This phrase-translation table was then used to create input sentences in which the L2 fragments are swapped for their L1 counterparts, effectively mimicking a fall-back to L1 in an L2 context. The full L2 sentence acts as reference sentence. Finally, to ensure all fragments are correct and sensible, a manual selection from this automatically generated corpus constituted the final trial set.

In our pilot study, such a data set, even without the manual selection stage, proved adequate to demonstrate the feasibility of translating L1 fragments in an L2 context (van Gompel and van den Bosch, 2014). One can, however, rightfully argue whether such data is sufficiently representative for the task and whether it would adequately cover instances where L2 language learners might experience difficulties and be inclined to fall back to L1. We therefore created a more representative test set for the task.

The actual test set conforms to much more stringent constraints and was composed entirely by hand from a wide variety of written sources. Amongst these sources are study books and grammar books for language learners, short bilingual on-line stories aimed at language learners, gap-exercises and cloze tests, and contemporary written resources such as newspapers, novels, and Wikipedia. We aimed for actual learner corpora, but finding suitable learner corpora with sufficient data proved hard. For German we could use the the Merlin corpus (Abel et al., 2013). In example (a) we see a real example of a fragment in a fall-back language in an L2 context from the Merlin corpus.

(a) **Input:** Das Klima hier ist **Tropical** und wir haben fast keinen Winter
Reference: Das Klima hier ist **tropisch** und wir haben fast keinen Winter.

For various sources bilingual data was available. For the ones that were monolingual (L2) we resorted to manual translation. To ensure our translations were correct, these were later independently verified, and where necessary corrected by native speakers.

A large portion of the test set comes from off-line resources because we wanted to make sure that a substantial portion of the test set could not be found verbatim on-line. This was done to prevent systems from solving the actual problem by just attempting to just look up the sources through the available context information.

Note that in general we aimed for the European varieties of the different languages. However, for English we did add the US spelling variants as alternatives. A complete list of all sources used in establishing the test set is available on our website[1].

We created a trial set and test set/gold standard of 500 sentence pairs per language pair. Due to the detection of some errors at a later stage, some of which were caused by the tokenisation process, we were forced to remove some sentences from the test set and found ourselves slightly below our aim for some of the language pairs. The test set was delivered in both tokenised[2] and untokenised form. The trial set was delivered only in tokenised form. Evaluation was conducted against the tokenised version, but our evaluation script was designed to be as lenient as possible regarding differences in tokenisation. We explicitly took cases into account where participant's tokenisers split contractions (such as Spanish "del" to "de" + "el"), whereas our tokeniser did not.

For a given input fragment, it may well be possible that there are multiple correct translations possible. In establishing our test set, we therefore paid special attention to adding alternatives. To ensure no alternatives were missed, all participant output was aggregated in one set, effectively anonymising the systems, and valid but previously missed alternatives were added to the gold standard.

4 Evaluation

Several metrics are available for automatic evaluation. First, we measure the absolute accuracy $a = c/n$, where c is the number of fragment translations from the system output that precisely match the corresponding fragments in the reference translation, and n is the total number of translatable fragments, including those for which no translation was found. We also introduce a word-based accuracy, which unlike the absolute accuracy gives some credits to mismatches that show partial overlap with the reference translation. It assigns a score according to the longest consecutive matching substring between output fragment and reference fragment and is computed as follows:

$$wac = \frac{|longestsubmatch(output, reference)|}{max(|output|, |reference|)} \quad (1)$$

The system with the highest word-based accuracy wins the competition. All matching is case-sensitive.

[1] https://github.com/proycon/semeval2014task5

[2] Using ucto, available at https://github.com/proycon/ucto

Systems may decide not to translate fragments if they cannot find a suitable translation. A recall metric simply measures the number of fragments for which the system generated a translation, regardless of whether that translation is correct or not, as a proportion of the total number of fragments.

In addition to these task-specific metrics, standard MT metrics such as BLEU, NIST, METEOR and error rates such as WER, PER and TER, are included in the evaluation script as well. Scores such as BLEU will generally be high (> 0.95) when computed on the full sentence, as a large portion of the sentence is already translated and only a specific fragment remains to be evaluated. Nevertheless, these generic metrics are proven in our pilot study to follow the same trend as the more task-specific evaluation metrics, and will be omitted in the result section for brevity.

It regularly occurs that multiple translations are possible. As stated, in the creation of the test set we have taken this into account by explicitly encoding valid alternatives. A match with any alternative in the reference counts as a valid match. For word accuracy, the highest word accuracy amongst all possible alternatives in the reference is taken. Likewise, participant system output may contain multiple alternatives as well, as we allowed two different types of runs, following the example of the Cross-Lingual Lexical Substitution and Cross-Lingual Word Sense Disambiguation tasks:

- **Best** - The system may only output one, its best, translation;
- **Out of Five** - The system may output up to five alternatives, effectively allowing 5 guesses. Only the best match is counted. This metric does *not* count how many of the five are valid.

Participants could submit up to three runs per language pair and evaluation type.

5 Participants

Six teams submitted systems, three of which participated for all language pairs. In alphabetic order, these are:

1. **CNRC** - Cyril Goutte, Michel Simard, Marine Carpuat - National Research Council Canada – *All language pairs*
2. **IUCL** - Alex Rudnick, Liu Can, Levi King, Sandra Kübler, Markus Dickinson - Indiana University (US) – *all language pairs*
3. **UEdin** - Eva Hasler - University of Edinburgh (UK) – *all language pairs except English-German*
4. **UNAL** - Sergio Jiménez, Emilio Silva - Universidad Nacional de Colombia – *English-Spanish*
5. **Sensible** - Liling Tan - Universität des Saarlandes (Germany) and Nanyang Technological University (Singapore) – *all language pairs*
6. **TeamZ** - Anubhav Gupta - Université de Franche-Comté (France) – *English-Spanish, English-German*

Participants implemented distinct methodologies and implementations. One obvious avenue of tackling the problem is through standard Statistical Machine Translation (SMT). The CNRC team takes a pure SMT approach with few modifications. They employ their own Portage decoder and directly send an L1 fragment in an L2 context, corresponding to a partial translation hypothesis with only one fragment left to decode, to their decoder (Goutte et al., 2014). The UEdin team applies a similar method using the Moses decoder, marking the L2 context so that the decoder leaves this context as is. In addition they add a context similarity feature for every phrase pair in the phrase translation table, which expresses topical similarity with the test context. In order to properly decode, the phrase table is filtered per test sentence (Hasler, 2014). The IUCL and UNAL teams do make use of the information from word alignments or phrase translation tables, but do not use a standard SMT decoder. The IUCL system combines various information sources in a log-linear model: phrase table, L2 Language Model, Multilingual Dictionary, and a dependency-based collocation model, although this latter source was not finished in time for the system submission (Rudnick et al., 2014). The UNAL system extracts syntactic features as a means to relate L1 fragments with L2 context to their L2 fragment translations, and uses memory-based classifiers to achieve this (Silva-Schlenker et al., 2014). The two systems on the lower end of the result spectrum use different techniques altogether. The Sensible team approaches the problem by attempting to emulate the manual post-editing process human translators employ to correct MT output (Tan et al., 2014), whereas TeamZ relies on Wiktionary as the sole source (Gupta, 2014).

6 Results

The results of the six participating teams can be viewed in consensed form in Table 1. This table shows the highest word accuracy achieved by the participants, in which multiple system runs have been aggregated. A ranking can quickly be distilled from this, as the best score is marked in bold. The system by the University of Edinburgh emerges as the clear winner of the task. The full results of the various system runs by the six participants are shown in Tables 2 and 3, two pages down, all three aforementioned evaluation metrics are reported there and the systems are sorted by word accuracy per language pair and evaluation type.

Team	**en-es**	oof	**en-de**	oof
CNRC	0.745	0.887	0.717	**0.868**
IUCL	0.720	0.847	**0.722**	0.857
UEdin	**0.827**	**0.949**	-	-
UNAL	0.809	0.880	-	-
Sensible	0.351	0.231	0.233	0.306
TeamZ	0.333	0.386	0.293	0.385
	fr-en	oof	**nl-en**	oof
CNRC	0.694	0.839	0.610	0.723
IUCL	0.682	0.800	0.679	0.753
UEdin	**0.824**	**0.939**	**0.692**	**0.811**
UNAL	-	-	-	-
Sensible	0.116	0.14	0.152	0.171
TeamZ	-	-	-	-

Table 1: Highest word accuracy per team, per language pair, and per evaluation type (out-of-five is include in the "oof" column). The best score in each column is marked in bold.

For the lowest-ranking participants, the score is negatively impacted by the low recall; their systems could not find translations for a large number of fragments.

Figures 1 (next page) and 2 (last page) show the results for the *best* evaluation type for each system run. Three bars are shown; from left to right these represent *accuracy* (blue), *word-accuracy* (green) and *recall* (red). Graphs for *out-of-five* evaluation were omitted for brevity, but tend to follow the same trend with scores that are somewhat higher. These scores can be viewed on the result website at `http://github.com/proycon/semeval2014task5/`. The result website also holds the system output and evaluation scripts with which all graphs and tables can be reproduced.

We observe that the best scoring team in the task (UEdin), as well as the CNRC team, both employ standard Statistical Machine Translation and achieve high results. From this we can conclude that standard SMT techniques are suitable for this task. Teams IUCL and UNAL achieve similarly good results, building on word and phrase alignment data as does SMT, yet not using a traditional SMT decoder. TeamZ and Sensible, the two systems ranked lowest do not rely on any techniques from SMT. To what extent the context-informed measures of the various participants are effective can not be judged from this comparison, but can only be assessed in comparison to their own baselines. For this we refer to the system papers of the participants.

7 Discussion

We did not specify any training data for the task. The advantage of this is that participants were free to build a wider variety of systems from various sources, rather than introducing a bias towards for instances statistical systems. The disadvantage, however, is that a comparison of the various systems does not yield conclusive results regarding the merit of their methodologies. Discrepancies might at least be partly due to differences in training data, as it is generally well understood in MT that more training data improves results. The baselines various participants describe in their system papers provide more insight to the merit of their approaches than a comparison between them.

In the creation of the test set, we aimed to mimic intermediate to high-level language learners. We also aimed at a fair distribution of different part-of-speech categories and phrasal length. The difficulty of the task differs between language pairs, though not intentionally so. We observe that the Dutch-English set is the hardest and the Spanish-English is the easiest in the task. One of the participants implicitly observes this through measurement of the number of Out-of-Vocabulary words (Goutte et al., 2014). This implies that when comparing system performance between different language pairs, one can not simply ascribe a lower result to a system having more difficulty with said

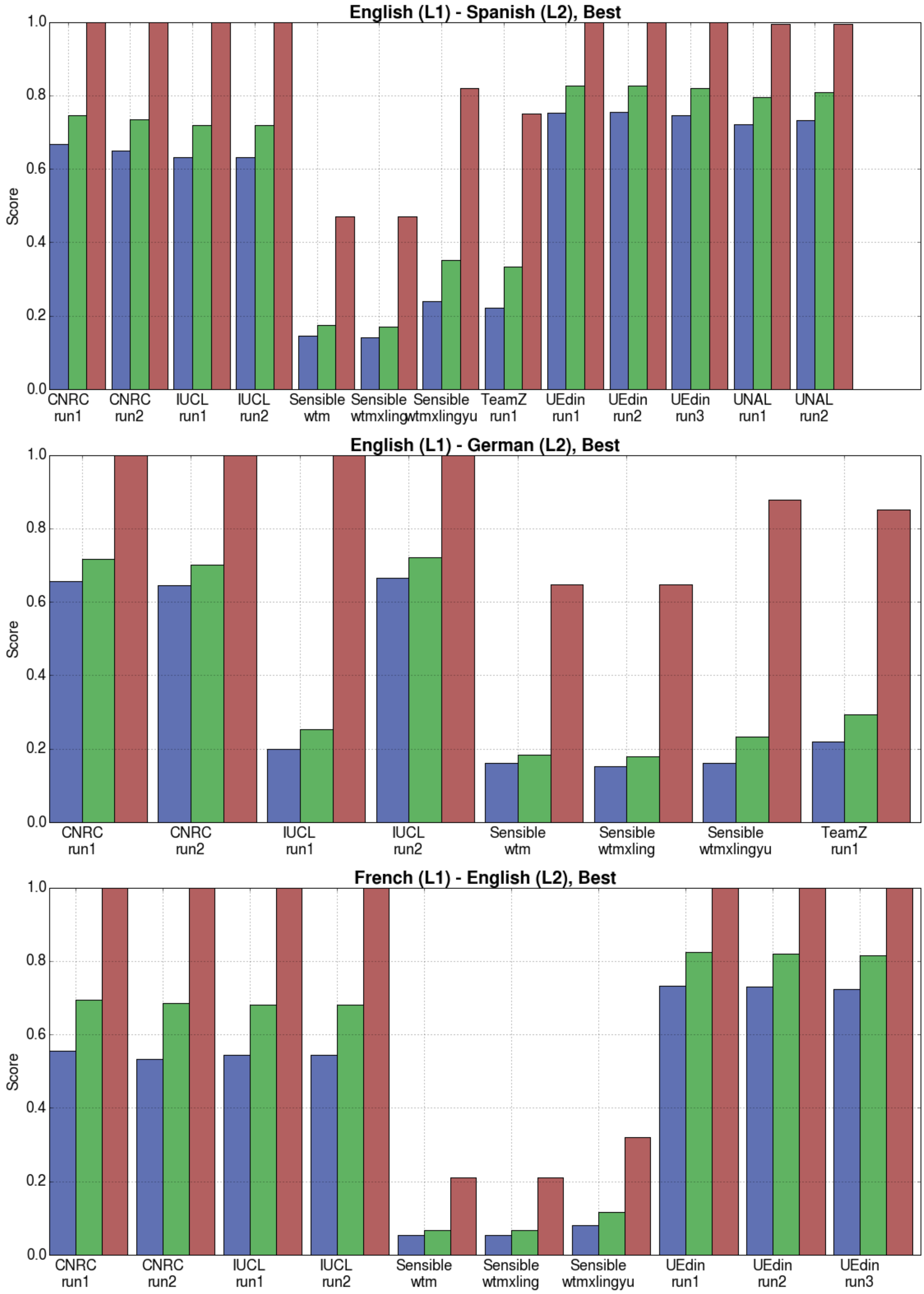

Figure 1: English to Spanish (top), English to German (middle) and French to English (bottom). The three bars, left-to-right, represent Accuracy (blue), Word Accuracy (green) and Recall (red).

System	Acc	W.Acc.	Recall
English-Spanish (best)			
UEdin-run2	0.755	0.827	1.0
UEdin-run1	0.753	0.827	1.0
UEdin-run3	0.745	0.82	1.0
UNAL-run2	0.733	0.809	0.994
UNAL-run1	0.721	0.794	0.994
CNRC-run1	0.667	0.745	1.0
CNRC-run2	0.651	0.735	1.0
IUCL-run1	0.633	0.72	1.0
IUCL-run2	0.633	0.72	1.0
Sensible-wtmxlingyu	0.239	0.351	0.819
TeamZ-run1	0.223	0.333	0.751
Sensible-wtm	0.145	0.175	0.470
Sensible-wtmxling	0.141	0.171	0.470
English-Spanish (out-of-five)			
UEdin-run3	0.928	0.949	1.0
UEdin-run1	0.924	0.946	1.0
UEdin-run2	0.92	0.944	1.0
CNRC-run1	0.843	0.887	1.0
CNRC-run2	0.837	0.884	1.0
UNAL-run1	0.823	0.88	0.994
IUCL-run1	0.781	0.847	1.0
IUCL-run2	0.781	0.847	1.0
Sensible-wtmxlingyu	0.263	0.416	0.819
TeamZ-run1	0.277	0.386	0.751
Sensible-wtm	0.173	0.231	0.470
Sensible-wtmxling	0.169	0.228	0.470
English-German (best)			
IUCL-run2	0.665	0.722	1.0
CNRC-run1	0.657	0.717	1.0
CNRC-run2	0.645	0.702	1.0
TeamZ-run1	0.218	0.293	0.852
IUCL-run1	0.198	0.252	1.0
Sensible-wtmxlingyu	0.162	0.233	0.878
Sensible-wtm	0.16	0.184	0.647
Sensible-wtmxling	0.152	0.178	0.647
English-German (out-of-five)			
CNRC-run1	0.834	0.868	1.0
CNRC-run2	0.828	0.865	1.0
IUCL-run2	0.806	0.857	1.0
TeamZ-run1	0.307	0.385	0.852
IUCL-run1	0.228	0.317	1.0
Sensible-wtmxlingyu	0.18	0.306	0.878
Sensible-wtm	0.182	0.256	0.647
Sensible-wtmxling	0.174	0.25	0.647

Table 2: Full results for English-Spanish and English-German.

System	Acc	W.Acc.	Recall
French-English (best)			
UEdin-run1	0.733	0.824	1.0
UEdin-run2	0.731	0.821	1.0
UEdin-run3	0.723	0.816	1.0
CNRC-run1	0.556	0.694	1.0
CNRC-run2	0.533	0.686	1.0
IUCL-run1	0.545	0.682	1.0
IUCL-run2	0.545	0.682	1.0
Sensible-wtmxlingyu	0.081	0.116	0.321
Sensible-wtm	0.055	0.067	0.210
Sensible-wtmxling	0.055	0.067	0.210
French-English (out-of-five)			
UEdin-run2	0.909	0.939	1.0
UEdin-run1	0.905	0.938	1.0
UEdin-run3	0.907	0.937	1.0
CNRC-run1	0.739	0.839	1.0
CNRC-run2	0.731	0.834	1.0
IUCL-run1	0.691	0.8	1.0
IUCL-run2	0.691	0.8	1.0
Sensible-wtmxlingyu	0.085	0.14	0.321
Sensible-wtmxling	0.061	0.09	0.210
Sensible-wtm	0.061	0.089	0.210
Dutch-English (best)			
UEdin-run1	0.575	0.692	1.0
UEdin-run2	0.567	0.688	1.0
UEdin-run3	0.565	0.688	1.0
IUCL-run1	0.544	0.679	1.0
IUCL-run2	0.544	0.679	1.0
CNRC-run1	0.45	0.61	1.0
CNRC-run2	0.444	0.609	1.0
Sensible-wtmxlingyu	0.115	0.152	0.335
Sensible-wtm	0.092	0.099	0.214
Sensible-wtmxling	0.088	0.095	0.214
Dutch-English (out-of-five)			
UEdin-run1	0.733	0.811	1.0
UEdin-run3	0.727	0.808	1.0
UEdin-run2	0.725	0.808	1.0
IUCL-run1	0.634	0.753	1.0
IUCL-run2	0.634	0.753	1.0
CNRC-run1	0.606	0.723	1.0
CNRC-run2	0.602	0.721	1.0
Sensible-wtmxlingyu	0.123	0.171	0.335
Sensible-wtm	0.099	0.115	0.214
Sensible-wtmxling	0.096	0.112	0.214

Table 3: Full results for French-English and Dutch-English.

language pair. This could rather be an intrinsic property of the test set or the distance between the languages.

Distance in syntactic structure between languages also defines the limits of this task. During composition of the test set it became clear that backing off to L1 was not always possible when syntax diverged to much. An example of this is separable verbs in Dutch and German. Consider the German sentence *"Er **ruft** seine Mutter **an**"* (translation: *"He **calls** his mother"*). Imagine a German language learner wanting to compose such a sentence but wanting to fall back to English for the verb *"to call"*, which would translate to German as *"anrufen"*. The possible input sentence may still be easy to construe: *"Er **calls** seine Mutter"*, but the solution to this problem would require insertion at two different points, whereas the task currently only deals with a substitution of a single fragment. The reverse is arguably even more complex and may stray too far from what a language learner may do. Consider an English language learner wanting to fall back to her native German, struggling with the English translation for *"anrufen"*. She may compose a sentence such as *"He **ruft** his mother **an**"*, which would require translating two dependent fragments into one.

We already have interesting examples in the gold standard, such as example (b), showing syntactic word-order changes confined to a single fragment.

(b) **Input:** I always wanted **iemand te zijn** , but now I realize I should have been more specific.
Reference: I always wanted **to be somebody** , but now I realize I should have been more specific.
Participant output (aggregated): to be a person; it to be; someone to his; to be somebody; person to be; someone to; someone to be; to be anybody; to anyone; to be someone; a person to have any; to be someone else

Another question we can ask, but have not investigated, is whether a language learner would insert the proper morphosyntactic form of an L1 word given the L2 context, or whether she may be inclined to fall back to a normal form such as an infinitive. Especially in the above case of separable verbs someone may be more inclined to circumvent the double fragments and provide the input: *"He **anrufen** his mother"*, but in simpler cases the same issue arises as well. Consider an English learner falling back to her native Croatian, a Slavic language which heavily declines nouns. If she did not know the English word *"book"* and wanted to write *"He gave the book to him"*, she could use either the Croatian word *"knjigu"* in its accusative declension or fall back to the normal form *"knjiga"*. A proper writing assistant system would have to account for both options.

We can analyse which of the sentences in the test data participants struggled with most. First we look at the number of sentences that produce an average word accuracy of zero, measured per sentence over all systems and runs in the out-of-five metric. This means no participant was close to the correct output. There were 6 such sentences in English-Spanish, 17 in English-German, 6 in French-English, and 32 in Dutch-English.

A particularly difficult context from the Spanish set is when a subjunctive verb form was required, but an indicative verb form was submitted by the systems, such as in the sentence: *"Espero que los frenos del coche **funcionen** bien."*. Though this may be deduced from context (the word *"Espero"*, expressing hope yet doubt, being key here), it is often subtle and hard to capture. Another problematic case that recurs in the German and Dutch data sets is compound nouns. The English fragment *"work motivation"* should translate into the German compound *"Arbeitsmotivation"* or *"Arbeitsmoral"*, yet participants were not able to find the actual compound noun. Beside compound nouns, other less frequent multi-word expressions are also amongst the difficult cases. Sparsity or complete absence in training data of these expressions is why systems struggle here.

Another point of discussion is the fact that we enriched the test set by adding previously unavailable alternative translations from an aggregated pool of system output. This might draw criticism for possibly introducing a bias, also considering the fact that the decision to include a particular alternative for a given context is not always straightforward and at times subjective. We, however, contend that this is the best way to ensure that valid system output is not discarded and reduce the number of false negatives. The effect of this measure has been an increase in (word) accuracy for all systems, without significant impact on ranking.

8 Conclusion

In this SemEval task we showed that systems can translate L1 fragments in an L2 context, a task that finds application in computer-assisted translation and computer-assisted language learning. The localised translation of a fragment in a cross-lingual context makes it a novel task in the field. Though the task has its limits, we argue for its practical application in a language-learning setting: as a writing assistant and dictionary replacement. Six contestants participated in the task, and used an ensemble of techniques from Statistical Machine Translation and Word Sense Disambiguation. Most of the task organizers' time went into manually establishing a gold standard based on a wide variety of sources, most aimed at language learners, for each of the four language pairs in the task. We have been positively surprised by the good results of the highest ranking systems.

9 Acknowledgements

We would like to thank Andreu van Hooft and Sarah Schulz for their manual correction work, and Sean Banville, Geert Joris, Bernard De Clerck, Rogier Crijns, Adriane Boyd, Detmar Meurers, Guillermo Sanz Gallego and Nils Smeuninx for helping us with the data collection.

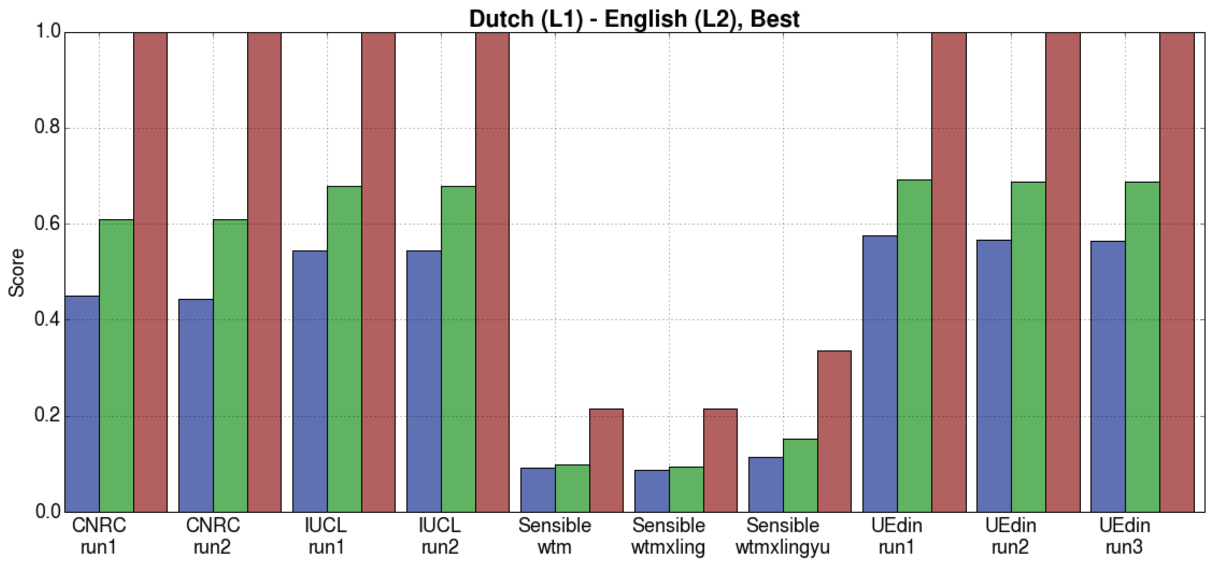

Figure 2: Dutch to English.

References

Andrea Abel, Lionel Nicolas, Jirka Hana, Barbora Štindlová, Katrin Wisniewski, Claudia Woldt, Detmar Meurers, and Serhiy Bykh. 2013. A trilingual learner corpus illustrating european reference levels. In *Proceedings of the Learner Corpus Research Conference*, Bergen, Norway, 27-29 September.

Cyril Goutte, Michel Simard, and Marine Carpuat. 2014. CNRC-TMT: Second language writing assistant system description. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Anubhav Gupta. 2014. Team Z: Wiktionary as L2 writing assistant. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Eva Hasler. 2014. UEdin: Translating L1 phrases in L2 context using context-sensitive smt. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ohttp://www.aclweb.org/anthology/P/P07/P07 2045ndrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *In Proceedings of the Machine Translation Summit X ([MT]'05).*, pages 79–86.

Els Lefever and Veronique Hoste. 2013. SemEval-2013 Task 10: Cross-Lingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics*.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval 2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Franz Josef Och and Hermann Ney. 2000. Giza++: Training of statistical translation models. Technical report, RWTH Aachen, University of Technology.

Alex Rudnick, Levi King, Can Liu, Markus Dickinson, and Sandra Kübler. 2014. IUCL: Combining information sources for semeval task 5. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Emilio Silva-Schlenker, Sergio Jimenez, and Julia Baquero. 2014. UNAL-NLP: Cross-lingual phrase sense disambiguation with syntactic dependency trees. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Liling Tan, Anne Schumann, José Martinez, and Francis Bond. 2014. Sensible: L2 translation assistance by emulating the manual post-editing process. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland.

Maarten van Gompel and Antal van den Bosch. 2014. Translation assistance by translation of L1 fragments in an L2 context. In *To appear in Proceedings of ACL 2014*.