WAGS: A Beautiful English-Italian Benchmark Supporting Word Alignment Evaluation on Rare Words

Luisa Bentivogli[1], Mauro Cettolo[1], M. Amin Farajian[1,2], Marcello Federico[1]
[1]Fondazione Bruno Kessler, Via Sommarive 18, Povo - Trento, Italy
[2]University of Trento, Via Sommarive, 5, Povo - Trento, Italy
{bentivo,cettolo,farajian,federico}@fbk.eu

Abstract

This paper presents WAGS (Word Alignment Gold Standard), a novel benchmark which allows extensive evaluation of WA tools on out-of-vocabulary (OOV) and rare words. WAGS is a subset of the Common Test section of the Europarl English-Italian parallel corpus, and is specifically tailored to OOV and rare words. WAGS is composed of 6,715 sentence pairs containing 11,958 occurrences of OOV and rare words up to frequency 15 in the Europarl Training set (5,080 English words and 6,878 Italian words), representing almost 3% of the whole text. Since WAGS is focused on OOV/rare words, manual alignments are provided for these words only, and not for the whole sentences. Two off-the-shelf word aligners have been evaluated on WAGS, and results have been compared to those obtained on an existing benchmark tailored to full text alignment. The results obtained confirm that WAGS is a valuable resource, which allows a statistically sound evaluation of WA systems' performance on OOV and rare words, as well as extensive data analyses. WAGS is publicly released under a Creative Commons Attribution license.

Keywords: Word Alignment, Gold Standard dataset, evaluation, out-of-vocabulary words, rare words

1. Introduction

The task of Word Alignment (WA) consists of finding the correspondence between words that are translation of each other in a bilingual sentence pair (Brown et al., 1990). WA is a basic component of Statistical Machine Translation (Och and Ney, 2004; Fraser and Marcu, 2007), but also other applications rely on WA, such as extraction of bilingual lexica (Smadja et al., 1996), word sense disambiguation (Diab and Resnik, 2002), projection of linguistic information between languages (Yarowsky and Ngai, 2001; Kuhn, 2004; Bentivogli and Pianta, 2005).

WA gold standards represent a crucial resource to evaluate and analyse WA systems' performance, and nowadays various benchmarks for different language pairs are available (Melamed, 1998; Och and Ney, 2000; Mihalcea and Pedersen, 2003; Lambert et al., 2005; Martin et al., 2005; Kruijff-Korbayová et al., 2006; Graça et al., 2008; Macken, 2010; Holmqvist and Ahrenberg, 2011). Besides the languages addressed, existing benchmarks differ in various respects – also depending on the final application to be evaluated – such as the parallel data used, the annotation scheme adopted (and related guidelines), the selection of words to be manually aligned. Regarding this latter issue, two main approaches were followed in previous works: *full text* alignment, where all words in the text are manually aligned, and *sample word* alignment, where a set of test words are selected and only those words are manually aligned (Véronis and Langlais, 2000; Merkel, 1999; Ahrenberg et al., 2002).

One of the most challenging issues for current state-of-the-art word aligners is that they show poor generalization capability and are prone to errors when infrequent or unknown words (with respect to the training data) occur in new sentence pairs to be aligned (Farajian et al., 2014). Thus, WA research would highly benefit from gold standard data specifically tailored to assess WA systems on this issue. However, to our knowledge, none of the available WA benchmarks specifically focuses on the problem of out-of-vocabulary (OOV) and rare words.

The main contribution of our work is to provide the research community with WAGS (Word Alignment Gold Standard), a novel benchmark which allows extensive evaluation of WA tools on OOV and rare words. WAGS is a subset of the Common Test section of the Europarl English-Italian parallel corpus (Koehn et al., 2003; Koehn, 2005), and is specifically tailored to OOV and rare words. WAGS is composed of 6,715 sentence pairs containing 11,958 occurrences of OOV and rare words up to frequency 15 in the Europarl Training set (5,080 in the English side and 6,878 on the Italian side). These words represent about the 3% of the full text (specifically, 2.3% of the English side, and 3.2% of the Italian side), to be compared to the 0.7% measured on the full Common Test set or similar values of other standard Europarl-based corpora used for MT evaluation. Since WAGS is focused on OOV/rare words, manual alignments are provided for these words only, following the so-called sample word alignment method. The large size of our reference collection makes it a valuable resource, which allows a statistically sound evaluation of WA systems' performance on OOV and rare words, as well as extensive data analyses aimed at shading light on this crucial aspect for WA and – more generally – machine translation.

WAGS is publicly released under a Creative Commons Attribution license (CC BY 4.0) and is available at:

`http://hlt-mt.fbk.eu/technologies/wags`

In addition to the gold standard data, the release includes the annotation guidelines and an evaluation package that allows to compute Alignment Error Rate (AER) on customizable subsets of WAGS links, for example those aligning only OOV words.

In the following, we describe the characteristics of WAGS and provide results from relevant WA state-of-the-art technology, namely fast_align (Dyer et al., 2013) – a variant of IBM model 2 – and IBM model 4 as implemented in mgiza++ (Gao and Vogel, 2008).

2. Dataset Description

To create WAGS, we used the publicly available Europarl parallel corpus[1] (Koehn, 2005), which contains the proceedings of the European Parliament in the various official languages. A Common Test set, made of the texts from the 4th quarter of year 2000, was defined to be used for machine translation evaluation (Koehn et al., 2003). Table 1 shows some statistics about the Italian and English portions of Europarl v7 release. WAGS is a selection from the Common Test set, realized as described below.

	#seg	#Ita tokens	#Eng tokens
Training	1,908,966	54,848,640	55,141,541
Common Test	42,753	1,224,178	1,266, 968

Table 1: Europarl v7 statistics: number of segments and Italian/English tokens in Training and Common Test sets.

2.1. Data selection

The length of segments in the Europarl Common Test set ranges from one (single word segments) to more than two hundreds. It is a matter of fact that the automatic WA of either too short or too long segments can be overly easy/hard, respectively. In order to devise an effective benchmark for practical use, the segment pairs of the Common Test set of length falling in the 5% tails were not considered for the selection. The remaining segments include 8 to 60 tokens in the Italian side and 8 to 62 tokens in the English side. Figure 1 shows segment length distributions before and after the discarding process.

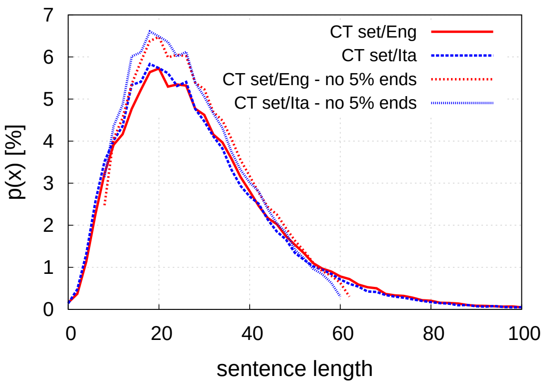


Figure 1: Distribution of lengths of English and Italian segments in the whole Common Test (CT) set before and after the 5% tails removal.

After having also filtered out pairs with anomalous length ratio (which are typically the result of wrong sentence alignment), we selected all and only pairs which include in either the Italian or English side at least an OOV word or a F[1,10] word.[2] In the resulting dataset, all F[0,10] words were manually aligned, as well as F[11,15] words. The F[11,15] words included in the dataset are not all those contained in the Common Test set. We manually aligned them in view of a future extension of WAGS with the segment pairs containing the remaining F[11,15] words. Even though representing a subset of the F[11,15] class, the gold alignments contained in the current version of WAGS constitute useful additional evaluation data. The statistics of the resulting dataset are reported in Table 2.

	Ita	Eng
Segments	6,715	
Total words	212,934	219,454
F[0,15]	6,878	5,080

Table 2: WAGS size statistics: number of segment pairs and tokens in the Italian/English sides.

Figure 2 plots the coverage rates of English F[`N`] words, `N`= $0 \ldots 15$, in WAGS, in the whole Common Test set from which WAGS was selected, and in the Europarl test set proposed in the MT shared task of WMT 2008 (`tst2008`), which should be a good representative of the whole Common Test set. Indeed, since `tst2008` is a random sample of the Common Test set, their plots are indistinguishable. On the contrary, the tailoring of our benchmark towards OOV/rare words is remarkable: in WAGS, the rate of F[`N`] words for any `N`= $0 \ldots 10$ is 4 to 5 times higher than in the Common Test set. The difference reduces for words with higher frequency since WAGS contains only a subset of them. Globally, F[0,15] English words cover 2.3% of WAGS, while in the other two considered sets they represent just 0.6%. Similar figures hold for the Italian side.

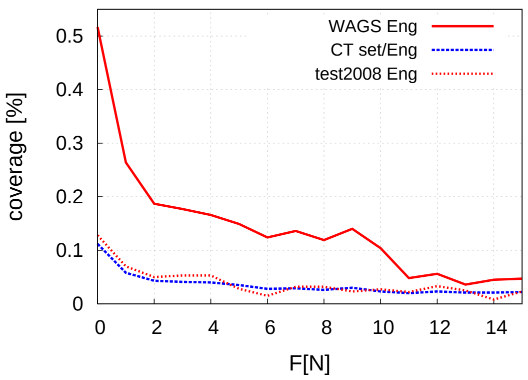


Figure 2: Percentage of the English side of WAGS, of the whole Common Test (CT) set and of the Europarl `test2008` covered by F[0] to F[15] words.

[1] `www.statmt.org/europarl/` (accessed March, 2016)

[2] We denote by `F[N,M]` the set of words with frequency between N and M in the training set; F[N] and F[N,] are shortcuts for F[N,N] and [N,∞], respectively. Note that F[0] is the set of OOVs.

2.2. Manual alignment

For the manual alignment task, detailed guidelines were produced. Existing guidelines (Lambert et al., 2005; Graça et al., 2008) were adopted so to allow proper comparison with other word alignment benchmarks, but were modified and augmented where necessary to be compliant with our specific alignment task and languages.[3]

For each sentence pair, annotators were not asked to align all the words, but only those selected as OOV and rare words. In order to ensure alignment completeness – as well as consistency with respect to a full text alignment perspective – annotators were asked to verify if the word to be aligned or its translation belonged to a complex lexical unit (*i.e.* a lexical unit composed of a group of words). If so, they had to align the whole lexical units, thus creating one-to-many/many-to-one as well as many-to-many links. For instance, in Example 1[4] the Italian verb "domanderete" was aligned to its corresponding translation "you will ask", while in Example 2 the Italian noun "archi" was part of a complex lexical unit "quartetto d'archi" which was aligned as a whole to "string quartet".

Example 1:

- Ita: allora, mi **domanderete** e so che mi verrà chiesto....
- Eng: you will therefore ask me, and I fully expect to be questioned on this...

Example 2:

- Ita: magari è una coppa di champagne e un quartetto d' **archi**....
- Eng: perhaps it is a glass of champagne and a string quartet...

Furthermore, following the distinction introduced in (Och and Ney, 2000), both S(ure) links and P(ossible) links were allowed. While S-links represent unambiguous alignments, P-links represent alignments that might or might not exist. P-links were used especially for free translations and function words that do not have an exact counterpart in the other language. For instance, in Example 1 the Italian counterpart of "you" (*i.e.* "voi") is missing, so in principle "you" could be left unaligned. However, since the information about grammatical person is present in the verb, "you" was aligned with a P-link to "domanderete".

Finally, words that did not have a correspondence in the other language were left unlinked, following the so-called "*no-null-align*" mode.

To ensure data quality, the whole dataset was annotated independently by two translators. At the end of each working day, annotators examined together all disagreements, with the double purpose of *(i)* fixing annotation errors that may have occurred, and *(ii)* discussing problematic cases – without necessarily trying to reach an agreement– in order to progressively refine the guidelines, thus making the annotation process more consistent. In this way, after the "reconciliation" phase, only real disagreement between annotators was left in the dataset.

Once the double annotation was collected for the whole dataset, we adopted the following strategy to produce the final gold standard alignment. First, an additional adjudication phase was carried out by a third judge, who solved the disagreements for which s/he found an appropriate solution. Example 3 shows a sentence pair where the Italian word "immemori" was to be aligned. The word is part of an expression ("da tempi immemori", Eng. "since time immemorial") which has a rather free counterpart in the English side; this yielded the two annotators to behave differently: one annotator created a P-link between the expressions "da tempi immemori" and "since its inception", while the other left the word unaligned. In this case the adjudicator selected the no-link option for the gold standard.

Example 3:

- Ita: ...un'aspirazione di questo Parlamento da tempi quasi **immemori**.
- Eng: ...an aspiration that has been held by Parliament almost since its inception.

For the disagreements that the adjudicator did not solve, we followed the common practice adopted in other available word alignment benchmarks to cope with alignment ambiguity, namely all links in disagreement were included in the final reference alignment as P-links. In Example 4, both annotators linked "orfana" ("orphan" in English) and "deprived", but one labeled it as S-link while the other as P-link. Since the adjudicator did not prefer one solution to the other, that link was labeled as Possible in the gold standard.

Example 4:

- Ita: ...questa Unione è stata un po' **orfana**.
- Eng: ...the Union was somewhat deprived.

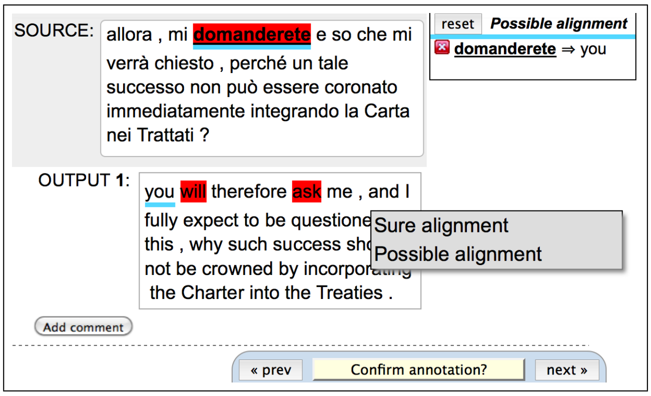
Annotations were accomplished using the MT-EQuAl toolkit[5] (Girardi et al., 2014). In addition to the traditional matrix-based alignment, MT-EQuAl allows a more user friendly text-based alignment procedure, where mouse clicks on words are used directly to establish alignment links. This alignment method was particularly useful in our task, since annotators were presented with long sentences and only few words were to be annotated. A screenshot of the alignment interface is presented in Figure 3, which shows the alignment of Example 1 above. In the figure, the P-link between "you" and "domanderete" has already been created: the words are underlined in the text (light blue) and the alignment is listed in the right part of the interface. Instead, the S-link between "domanderete" and "will ask"

[3] The complete alignment guidelines are released together with the gold standard data.

[4] In all the examples presented in the paper, OOV/rare words to be aligned are marked in bold, while aligned words are underlined.

[5] `github.com/hltfbk/mt-equal` (accessed March, 2016)

is being created: words were selected by clicking on them, and the menu for choosing between S-link and P-link (light grey box) was activated by right-clicking. Once the alignment type is chosen, the links are saved, the involved words are underlined, and the new alignment appears in the corresponding list on the right.

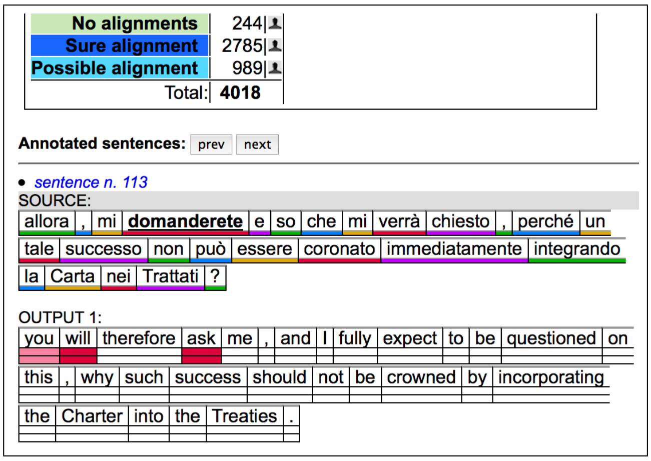


SOURCE: allora , mi domanderete e so che mi verrà chiesto , perché un tale successo non può essere coronato immediatamente integrando la Carta nei Trattati ?

OUTPUT 1: you will therefore ask me , and I fully expect to be questione this , why such success sho not be crowned by incorporating the Charter into the Treaties .

Figure 3: MT-EQuAl alignment interface.

Figure 4 shows the inter-annotation agreement interface that was used for the reconciliation phase. Alignment statistics are reported at the top of the page, while human annotations are shown under the target sentence, one line for each annotator (two annotators in agreement, in the example). Alignments are visualized through colors: a fixed color is assigned to each word in the source sentence, while colors under the target sentence word(s) indicate the alignment to the corresponding source word(s). Bright colors indicate S-links and pale colors indicate P-links. In Figure 4 we can see that "domanderete" is assigned red, and in the target sentence "you" is marked with pale red (corresponding to a P-link to "domanderete"), and "will" and "ask" are marked with bright red (corresponding to two S-links to "domanderete"). Furthermore, if the mouse is hovered over a colored target word, the corresponding aligned source words are highlighted.



No alignments	244
Sure alignment	2785
Possible alignment	989
Total:	4018

Annotated sentences: prev next

- sentence n. 113

SOURCE:
allora , mi domanderete e so che mi verrà chiesto , perché un tale successo non può essere coronato immediatamente integrando la Carta nei Trattati ?

OUTPUT 1:
you will therefore ask me , and I fully expect to be questioned on this , why such success should not be crowned by incorporating the Charter into the Treaties .

Figure 4: MT-EQuAl agreement visualization interface.

2.3. Gold standard alignment statistics

WAGS contains a total of 32,987 links, among which 26,617 are S-links and 6,370 are P-links. It is important to note that since many-to-many links were created in presence of complex lexical units, the gold standard contains also links between words that are not OOV/rare (*e.g.* the link between "quartetto" and "quartet" pertaining to the alignment of the rare word "archi" in Example 2 presented in Section 2.2.). The number of these links is 14,598, out of which 84% are S-links (12,218).

However, since in this work we restrict the evaluation to only OOV/rare words, both the detailed statistics presented in Table 3 and the evaluation results given in Tables 8 and 9 refer to all and only the alignment links directly involving OOV/rare words.

Table 3 shows the distribution of the WAGS links with respect to the frequency ranges of Italian and English words. We can see that the total number of links directly involving OOV/rare words is 18,389, out of which 78% (14,399) are S-links. Furthermore, it is interesting to see that the links involving words that are OOV or rare in both languages, and therefore particularly hard to automatically hypothesize, are 2,655, a number that should allow statistically sound investigations. We also notice the high number of alignments involving frequent words (F[16,]). These figures are related to the alignment of complex lexical units, as it will be explained in Section 3.2..

Finally, the number of OOV/rare words that occur in one language but do not have a correspondence in the other, and were thus left unaligned, is 670 for Italian and 342 for English.

WAGS		Eng side			
		F[0]	F[1,15]	F[16,]	total
Ita side	F[0]	468	246	1,372	2,086
	F[1,15]	220	1,721	7,446	9,387
	F[16,]	1,399	5,517	–	6,916
	total	2,087	7,484	8,818	18,389

Table 3: Distribution of WAGS alignments linking F[0] (OOV) or F[1,15] words either on both sides or on just one side of the dataset.

To measure annotators' consistency and assess the actual ambiguity of the alignment task, Inter-Annotator Agreement between the two annotators was calculated after the reconciliation phase. We adopted the standard formula

$$AGR = \frac{2*I}{A1+A2}$$

where $A1$ and $A2$ are the sets of links created by the two annotators and I is the intersection of these sets. Table 4 presents agreement results calculated following different principles: counting common alignment points without taking into account their type (no S/P distinction), or considering in agreement only those common links marked with the same alignment type by both annotators (S+P); in addition, agreement on S-links and P-links is given.

As we can see, agreement is quite high, confirming the reliability of our alignment guidelines and the consistency of

Links	% Agreement
S	96.03
P	86.22
S+P	93.96
No S/P distinction	98.12

Table 4: Inter-annotator agreement.

annotations. These results are slightly higher but can be considered in line with those obtained in other annotation projects (Kruijff-Korbayová et al., 2006; Graça et al., 2008) where particular attention was paid to the creation of very detailed annotation guidelines able to give clear rules for particularly difficult or data-specific cases. Indeed, in those datasets full text alignment was performed, which involves the alignment of a much higher number of function words, which are the most difficult to align (Melamed, 1998). Differently, in WAGS OOV and rare words are typically content words, and a relevant number is represented by named entities whose corresponding translation is usually clear.

3. Evaluation

The reliability of results on WAGS was compared to that on an existing benchmark tailored to full text alignment, which is similar in size to usual WA benchmarks. Automatic alignments were computed on both benchmarks by means of "off-the-shelf" state-of-the-art word aligners, namely fast_align (Dyer et al., 2013), that implements a reparametrization of IBM model 2 for training, and mgiza++ (Gao and Vogel, 2008), set up for training IBM model 4. The models were estimated on training data only, not including the test set. Although the test set could be fairly included in the training data, being WA an unsupervised task, we decided not to include it based on some considerations. On the one hand, we want to study the behavior of WA systems on those WAGS words which are actually OOV or rarely observed in the training text; on the other hand, this setting reflects real scenarios, such as Computer Assisted Translation integrating adaptive MT systems, where new sentence pairs are to be aligned with already trained word aligners.

The full alignment benchmark (FAB) used in our evaluation is a subset of the Italian/English JRC-legal corpus (Steinberger et al., 2006), manually aligned as described in (Farajian et al., 2014). It consists of 200 parallel sentences, with about 7,000 tokens per side. The total number of reference alignments is 7,380, the distribution of which is presented in Table 5.

FAB		Eng side			
		F[0]	F[1,15]	F[16,]	total
Ita side	F[0]	37	9	31	77
	F[1,15]	17	34	103	154
	F[16,]	13	85	7,051	7149
	total	67	128	7,185	7380

Table 5: Distribution of FAB alignment links.

For training puposes, we used the same subset of JRC-legal described in (Farajian et al., 2014), which is disjoint from FAB and includes about a million sentences and 20 million words per side.

3.1. Results

The results obtained by running fast_align[6] and mgiza++[7] on FAB are presented in Tables 6 and 7, in terms of AER as defined in (Martin et al., 2005).

fast_align AER on FAB		Eng side F[0]	F[1,15]	F[16,]
Ita side	F[0]	42.31	80.00	83.87
	F[1,15]	65.22	22.81	63.95
	F[16,]	89.74	44.52	15.14

Table 6: Performance of fast_align on FAB. Percentage AERs are provided for different frequency classes.

mgiza++ AER on FAB		Eng side F[0]	F[1,15]	F[16,]
Ita side	F[0]	68.18	50.00	69.57
	F[1,15]	36.00	8.24	27.04
	F[16,]	45.00	17.86	11.15

Table 7: Performance of mgiza++ on FAB. Percentage AERs are provided for different frequency classes.

As evident from Table 5, the great majority of links in FAB involves frequent words (F[16,]) on both sides. On them, mgiza++ significantly outperforms fast_align, by 4 AER absolute points, 36% relative (11.15 vs. 15.14). On the other classes of links, the behaviour of the two aligners is similar, and sometimes counterintuitive. In fact, there are cases in which alignments are harder to hypothesize as the frequency of linked words increases. For example, mgiza++ alignes F[1,15]-F[1,15] words better than F[1,15]-F[16,] words (AER=8.24 vs. 27.04/17.86), and also even better than F[16,]-F[16,] words (AER=11.15). This issue has an explanation and is discussed in Section 3.2.

Surprisingly, on links between OOV words, fast_align performs better than mgiza++ (42.31 vs. 68.18). Actually, the number of reference F[0]-F[0] links is 37 (Table 5): fast_align generates 15 links between OOVs (all correct), mgiza++ only 7 (again all correct): it is evident that with so small numbers the statistical significance of outcomes is questionable. We will come back on this after the presentation of experiments on WAGS.

Experiments on WAGS followed the same scheme used for FAB, namely estimation of fast_align and mgiza++ models on training data only (Table 1). AER figures, reported in Tables 8 and 9, are presented grouping the words according to their frequency in the training text.

A generally higher performance of mgiza++ with respect to fast_align is measured on WAGS as well. Again, the issue

[6] Options: "-I 10 -d -o -v".

[7] Number of iterations on models: 5 on IBM models 1 and HMM, 3 on IBM models 3 and 4.

fast_align		Eng side		
AER on WAGS		F[0]	F[1,15]	F[16,]
Ita side	F[0]	59.40	57.60	78.92
	F[1,15]	68.92	13.58	59.85
	F[16,]	77.03	61.51	–

Table 8: Performance of fast_align on WAGS. Percentage AERs are provided for different frequency classes.

mgiza++		Eng side		
AER on WAGS		F[0]	F[1,15]	F[16,]
Ita side	F[0]	41.90	27.64	54.39
	F[1,15]	23.93	6.86	38.83
	F[16,]	60.33	44.51	–

Table 9: Performance of mgiza++ on WAGS. Percentage AERs are provided for different frequency classes.

of poorer quality on F[16,] links than on F[1,15] links arises (see Section 3.2.). Differently from FAB, here mgiza++ performs better than fast_align also on links between OOV words (41.90 vs. 59.40). As a matter of fact, the number of F[0]-F[0] links in WAGS is one order of magnitude higher than in FAB (468 vs. 37). To further investigate the relevance of the size of the dataset, we computed the confidence interval (at $\alpha = 0.05$ level) of AER on OOV words by performing 100 random sampling with replacement (bootstrapping) of the automatically aligned test sentences, and computing the AER of each sample. The results are as follows:

fast_align	FAB	42.56 ± 1.73
mgiza++	FAB	69.13 ± 1.77
fast_align	WAGS	59.19 ± 0.53
mgiza++	WAGS	41.79 ± 0.46

As expected, confidence intervals on WAGS are much smaller than on FAB, three to four times. The relevant fact to be remarked is that the outcome on OOV words on FAB is refuted by the much more reliable experiments on WAGS. This is a shining example which shows the risk of relying on standard full alignment benchmarks for investigating the behaviour of WA models on OOV and rare words.

3.2. Analysis

The evaluation results presented in Section 3.1. highlighted apparently counterintuitive results. First, we noticed that the classes where frequent words F[16,] are involved show the worst AER scores. Furthermore, both WA systems perform particularly well on class F[1,15]-F[1,15]. In order to provide an explanation for these issues, further analyses were carried out on WAGS data, which – due to the large number of reference links – allow reliable generalizations.

Our analysis suggests that in both cases results are related to the alignment of complex lexical units, *i.e.* units composed of a group of words. As we saw in Section 2.2., aligning complex lexical units requires creating one-to-many/many-to-one as well as many-to-many links, which represent a known difficulty for WA models. In particular, given the definition of the AER metric used, the cases that most affect evaluation are those where the complex lexical units are aligned through S-links.

We calculated the average fertility (*i.e.* the number of links connecting a given word) of Italian words in WAGS considering S-links only, and it turned out that fertility of OOV/rare words amounts to 1.57, while fertility of frequent words is 3.13.

These figures demonstrate that the classes involving frequent words contain a large number of multiple alignments, and thus are more difficult cases for current WA systems.

In WAGS there is a high number of rare words which belong to complex lexical units, such as names of persons, organizations, places, titles of books, names of laws. We present some of the most relevant cases, where the complex lexical units are aligned through S-links.

A typical example is given by compound words which have a corresponding translation made of more than one word.

Example 5:

- Ita: ...la raccolta di capitali a livello paneuropeo...
- Eng: ...pan-European capital **fundraising**...

Here the rare word “fundraising” (F[1,15]) is connected to Italian through three S-links involving F[16,] words, but the WA system was able to align it only to “raccolta”.

Another very frequent case of multiple alignment involving mainly S-links is given by verbs: auxiliary verbs are part of a complex lexical unit together with the main verb; reflexive verbs are often realized or aligned to complex lexical units; also, since Italian is morphologically richer than English, a high number of rare verbs correspond to inflected forms which typically align with more than one English word.

Example 6:

- Ita: metto in guardia coloro che stanno **spaccando** l’ Europa...
- Eng: I wish to warn all those who are **ripping** the Union...

Here, the verbs “spaccando” and ”ripping” are rare (both F[1,15]), while “stanno” and “are” are not rare. Each rare verb is involved in two S-links (“spaccando”-“are”, “spaccando”-”ripping” and ”ripping”-“stanno”, ”ripping”-“spaccando”). In this case the WA system was able to align the rare verbs but did not find the links between the main verbs and the auxiliaries.

Finally, regarding the very high performance of WA systems on class F[1,15]-F[1,15], we see from the fertility counts that the OOV and rare words classes contain a larger number of one-to-one links, and are thus less problematic. In particular, this class is easier than the classes involving OOV words, since these words have been seen in the training set. Examples of one-to-one translations correctly aligned by the systems are compound names such as “scolaretto”-“schoolboy”, “amore-odio”-“love-hate”, or proper names. Furthermore, this class is rarely affected by the problem of many-to-many alignments. Indeed, as shown in Example 6, the correct alignment between “spaccando” and “ripping” falls in that class, whereas the errors made by the system, which was not able to align “spaccando”-“are” and ”ripping”-“stanno”, is counted in the F[16,] classes.

4. Conclusion

In this paper we presented WAGS, a new benchmark for word alignment tailored to OOV and rare words. WAGS is a subset of the Common Test section of the Europarl English-Italian parallel corpus; it is composed of 6,715 sentence pairs containing 213 thousand Italian words and 219 thousand English words; OOV and rare words up to frequency 15 cover about the 3% of the full text and are manually aligned.

WAGS is publicly released under a Creative Commons Attribution license (CC BY 4.0) and is available at:

`http://hlt-mt.fbk.eu/technologies/wags`

In addition to the gold standard data, the release includes the annotation guidelines and an evaluation package that allows to compute AER on subsets of WAGS alignments customizable on the basis of the frequency of the linked words in the training set.

As for future work, we plan to create a new release of WAGS containing reference links for all the remaining F[11,15] words not included in the current version.

5. Acknowledgments

We thank the translators who annotated the dataset: Ilaria Antonini, Anna Demin, Gaia Kessler, Daniele Mich, Giulia Roggi, Chiara Zoller. Our thanks also go to the anonymous reviewers for their helpful comments.

This work was supported by the MMT (Modern Machine Translation) project, financed by European Union’s Horizon 2020 research and innovation programme under grant agreement No 645487.

6. Bibliographical References

Ahrenberg, L., Andersson, M., and Merkel, M. (2002). A system for incremental and interactive word linking. In *Proc. of LREC*.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural Language Engineering*, 11(3):247–261.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proc. of ACL*, pp. 255–262.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL*.

Farajian, M. A., Bertoldi, N., and Federico, M. (2014). Online word alignment for online adaptive machine translation. In *Proc of EACL 2014 Workshop on Humans and Computer-assisted Translation (HaCaT)*, pp. 84–92, Gothenburg, Sweden.

Fraser, A. and Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Proc. of SETQA-NLP*, pp. 49–57.

Girardi, C., Bentivogli, L., Farajian, M. A., and Federico, M. (2014). MT-EQuAl: a toolkit for human assessment of machine translation output. In *Proc. of COLING: System Demonstrations*, pp. 120–123, Dublin, Ireland.

Graça, J., Pardal, J. P., Coheur, L., and Caseiro, D. (2008). Building a golden collection of parallel multi-language word alignment. In *Proc. of LREC*, Marrakech, Morocco.

Holmqvist, M. and Ahrenberg, L. (2011). A gold standard for English–Swedish word alignment. In *Proc. of NODALIDA*, pp. 106–113.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proc. of HLT-NAACL*, pp. 127–133, Edmonton, Canada.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit X*, pp. 79–86, Phuket, Thailand.

Kruijff-Korbayová, I., Chvátalová, K., and Postolache, O. (2006). Annotation guidelines for Czech-English word alignment. In *Proc. of LREC*.

Kuhn, J. (2004). Experiments in parallel-text based grammar induction. In *Proc. of ACL*, pp. 470–477, Barcelona, Spain.

Lambert, P., De Gispert, A., Banchs, R., and Mariño, J. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Macken, L. (2010). An annotation scheme and gold standard for Dutch–English word alignment. In *Proc. of LREC*, Valletta, Malta.

Martin, J., Mihalcea, R., and Pedersen, T. (2005). Word alignment for languages with scarce resources. In *Proc. of ACL Workshop on Building and Using Parallel Texts (ParaText)*, pp. 65–74.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The Blinker project. IRCS Technical report #98-07, University of Pennsylvania.

Merkel, M. (1999). Annotation style guide for the plug link annotator. Technical report, Linkoping University, March.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proc. of the HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 1–10.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proc. of ACL*.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–450.

Smadja, F., Hatzivassiloglou, V., and McKeown, K. R. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of LREC*, pp. 2142–2147, Genoa, Italy.

Véronis, J. and Langlais, P. (2000). Evaluation of parallel text alignment systems. In *Parallel Text Processing*, pp. 369–388. Springer.

Yarowsky, D. and Ngai, G. (2001). Inducing multilingua POS taggers and NP bracketers via robust projection across aligned corpora. In *Proc. of NAACL*.