A Graph-based Strategy to Streamline Translation Quality Assessments

Daniele Pighin Lluís Formiga Lluís Màrquez
Universitat Politècnica de Catalunya (UPC), Barcelona, 08034 Spain
daniele.pighin@gmail.com lluis.formiga@upc.edu lluism@lsi.upc.edu

Abstract

We present a detailed analysis of a graph-based annotation strategy that we employed to annotate a corpus of 11,292 real-world English to Spanish automatic translations with relative (ranking) and absolute (adequate/non-adequate) quality assessments. The proposed approach, inspired by previous work in Interactive Evolutionary Computation and Interactive Genetic Algorithms, results in a simpler and faster annotation process. We empirically compare the method against a traditional, explicit ranking approach, and show that the graph-based strategy: 1) is considerably faster, and 2) produces consistently more reliable annotations.

1 Introduction

The creation of an annotated corpus is a demanding, repetitive, time and energy consuming activity that, more often than not, causes conspicuous amounts of fatigue and frustration in the annotators. This frustration has been identified as one of the main factors that lead to noise and contradictions in the annotation process (Takagi, 2001; Zaenen, 2006). These aspects are especially evident in fields such as Interactive Evolutionary Computation (IEC), where unknown fitness functions or subject-centered evaluations make high-repetitive, labor-intensive evaluation tasks extremely common. Researchers in these areas are devoting considerable effort towards the definition of novel annotation protocols that can reduce the fatigue of the annotators and result in more accurate and reliable resources (Llorà et al., 2005b; Formiga et al., 2010; Alm, 2011).

In this paper, we select one of these annotation schemes, developed and successfully tested in the context of Active Interactive Genetic Algorithms (aiGAs) (Llorà et al., 2005b; Formiga et al., 2010), and employ it to manually annotate a corpus of more than 11,000 machine-translated texts with relative (rankings) and absolute (adequate/non adequate) quality assessments. The method is based on a explicit decomposition of the traditional approach for ranking annotation (i.e., a many-to-many comparison problem) into a set of pairwise ranking decisions, from which a full ranking of alternatives can be automatically reconstructed.

To evaluate the appropriateness of the approach for machine-translation quality assessments, we present two different experiments. In the first, we compare the effort and the accuracy of the traditional annotation scheme against those of the graph-based strategy for the ranking of noisy web-log automatic translations, and show that the latter, 1) yields a higher inter-annotator agreement, and 2) can be considerably faster, especially for text segments of short-medium length. In the second experiment, we show how the graph-based annotation framework could reproduce the human rankings for the WMT 2010 (Callison-Burch et al., 2010) evaluation campaign with 93% accuracy, while requiring considerably less annotation effort. Furthermore, we demonstrate a method to pipeline relative and absolute assessment annotations that largely reduces the annotation effort in the absolute assessments stage ($>$65%, in terms of number of annotated pairs).

The rest of the paper is structured as follows: in the next section we discuss relevant previous work in the related areas; next, in Section 3 we present in more detail the whole annotation framework; in Section 4 we present the outcome of the annotation experiment and the results of our comparative evaluation; finally, in Section 5 we draw our conclusions.

2 Related work

Recent surveys have pointed out the inappropriateness of building huge amounts of annotated resources based only on inter-annotator agreement as a criterion to establish the quality of the resource (Zaenen, 2006). Beyond inter-annotator agreement, some other indicators should also be considered to better understand the complexity of the task, such as label-internal divergence or inter-annotator variation (Formiga et al., 2010; Alm, 2011). This work is a strong motivation to study alternative ways of collecting user annotations in a reliable and unambiguous way.

Research in Interactive Evolutionary Computation (IEC) has already addressed concepts such as user fatigue or consistency in human-computer interaction tasks involving a large amount of repetitive effort (Takagi, 2001; Llorà et al., 2005b). This research has developed novel, robust algorithms (Active Interactive Genetic Algorithms, or aiGAs) that integrate interactive human input in effective ways (Llorà et al., 2005b) while providing indicators of the consistency and reliability of the annotated resources (Formiga et al., 2010). The suitability of these algorithms for NLP tasks was already suggested from a theoretical perspective by Alm (2011).

Concretely, aiGAs are based on obtaining a complete ranking of solutions from simple pairwise comparisons after building an ordered graph. Within these ordered graphs, cycles represent a clear contradiction of a proper partial ordering (Llorà et al., 2005a). This simple property is the pillar to erect a first quantitative measure of the consistency of the evaluations provided by the user. aiGAs modeling has already been found helpful in many audio, speech and language related fields such as speech synthesis (Alías et al., 2011), affect in text and speech (Alm and Llorà, 2006) or music applications (Yang and Chen, 2009). In this paper, we set up a large-scale annotation activity with the purpose of demonstrating the practicality of employing the aiGA-inspired partial ordering model for a complex and linguistically dense problem like automatic translation ranking.

The annotation of machine-translated texts with quality assessments is typically carried out as part of MT evaluation campaigns (Callison-Burch et al., 2010). In this context, human assessments are typically used to rank the competing systems or to measure the correlation between reference-based metrics (Papineni et al., 2002; Giménez and Màrquez, 2010) with human quality assessments. More recently, a renewed interest in confidence and quality estimation for MT (Specia et al., 2009; Banchs and Li, 2011) has triggered the development of ad-hoc corpora to be used for training supervised models of translation quality (Specia et al., 2010).

3 Annotation methodology

Our objective is to build a corpus of rankings *and* absolute quality annotations for alternative translations of the same source sentences. These two layers of annotation are complementary and useful in different ways, and they can be exploited to learn models of quality with different applications, i.e., to select among alternative translations or to discard unsatisfactory outputs.

We considered 1,882 real-world translation requests in English submitted to an online translation service. A professional translator corrected the most obvious typos, slang or chat abbreviations and provided reference translations into Spanish for all of them. We automatically translated the corrected sentences into Spanish with five different systems: one of them is a state-of-the-art phrase-based MT system based on Moses that we trained using Europarl (Koehn, 2005), newswire (Callison-Burch et al., 2010) and UN (Rafalovitch and Dale, 2009) parallel corpora; the remaining four systems are online commercial systems that we queried via their web APIs, namely *SDL/LanguageWeaver*[1], *Google Translate*[2], *Bing Translator*[3] and *Systran*[4].

As a quality criterion for the assessments we selected adequacy, i.e., the amount of information that is correctly conveyed by a translation. This choice is motivated by the results of a preliminary annotation in which we compared five different quality criteria (namely: adequacy, fluency, a combination of adequacy and fluency, post-editing effort and a subjective measure of translation "goodness") and showed

[1] http://www.freetranslation.com/
[2] http://translate.google.com
[3] http://www.microsofttranslator.com
[4] http://systransoft.com

that, for the kind of noisy data that we are considering, adequacy guarantees the highest self consistency and inter-annotator agreement (Pighin et al., 2012).

3.1 Ranking a set of translations through pairwise comparisons

The partial ordering approach only requires the annotators to focus on the differences between two specific translations at a time. Intuitively, this is a simpler problem than the many-to-many comparison that is required to fully rank a set of translations.

The annotators are presented one source sentence s and a pair of alternative translations h_i and h_j at a time. For each triplet $\langle s, h_i, h_j \rangle$ the annotators take a *ternary decision*, by marking the two translations as equivalent or expressing a preference for one over the other. In the annotation guidelines, the annotators are invited to opt for the "tie" decision in all cases in which, according to their judgment, they would be equally satisfied (or dissatisfied) with the adequacy of the two translations.

We ask the annotators to annotate just enough triplets $\langle s, h_i, h_j \rangle$ to build a connected graph $\mathcal{G} = \langle V, E \rangle$ from all the translation alternatives (6, in our setting, 5 automatic translations + 1 reference translation). The annotations are presented to the annotator following the algorithm:

1. Let $\mathcal{H}$ be a list containing sets of translations. Initialize $\mathcal{H}$ to the list whose elements are singletons containing each translation of the source sentence s, i.e. $\mathcal{H} \leftarrow [\{h_1\}, \ldots, \{h_6\}]$;
2. While $\mathrm{length}(\mathcal{H}) > 1$:
   (a) Initialize an empty list $\mathcal{N} = []$;
   (b) Shuffle $\mathcal{H}$;
   (c) For each $A \in \mathcal{H}$, if $\nexists A' \in \mathcal{N} \mid h_i \in A'$, $\forall h_i \in A$, then:
      i. Randomly select $B \in \mathcal{H}, B \neq A$;
      ii. Append $A \cup B$ to $\mathcal{N}$;
      iii. Ask the user to annotate the triplet $\langle s, h_i, h_j \rangle$, where $h_i \in A$ and $h_j \in B$;
   (d) $\mathcal{H} \leftarrow \mathcal{N}$.

In other words, we create a tournament-like bracket configuration of the translations, as exemplified in Figure 1(1). First, we ask the annotator to annotate three translation pairs, i.e., the pre-terminal nodes of the tournament tree. Then, we move on to the upper-level bracket in the tournament by combining a random translation from one of the lower brackets with one translation from the other lower bracket, until there is only one bracket left. By following this scheme, we can build a connected graph $\mathcal{G}$ of the six alternative translations for each source sentence by means of 6 pairwise comparisons. In general, given n alternative translations the number of comparisons necessary to implement the tournament is given by $2 \cdot \lceil \frac{n}{2} \rceil$, i.e., n comparisons if n is even and $n + 1$ if n is odd. An example of such graph is shown in Figure 1(2). Here, undirected edges represent a tie, whereas directed edges correspond to the cases in which the annotator preferred one translation over the other. This connected, partially undirected graph can easily be turned into a fully-directed graph $\mathcal{G}'$ by exploiting topological properties of the graph deriving from its construction. Given the partially ordered graph $\mathcal{G}$, where local ordering is the direct consequence of pairwise user decisions, aiGAs (Llorà et al., 2005b) produce the fully directed $\mathcal{G}'$ by exploiting the *dominance* of each vertex (Pareto, 1896), a procedure originally inspired by multiobjective evaluation (Coello, 2000; Deb et al., 2000). In a similar way, we define the dominance of a vertex v as the quantity

$$\hat{f}(v) = \delta(v) - \phi(v) \ , \tag{1}$$

where $\delta(v)$ is the number of vertices that v dominates, and $\phi(v)$ is the number of vertices by which v is dominated. The process of generating the global ordering also eliminates loops and inconsistencies from the graph (3) and collapses into a single vertex all the equivalent translations. In Figure 1(4) the dominance of each node is shown next to it. In this case, the best translation would be the one provided by LW, having $\hat{f}(h_{LW}) = 4$, while Bing's would be ranked last with $\hat{f}(h_{Bing}) = -4$.

It should be noted that explicitly ranking the six translation using a full-ranking annotation scheme would require an annotator to sort them and implicitly perform, in the best case, $\sim 6 \log_2 6 = 15.5$ pairwise comparisons. That is, the tournament-based approach theoretically reduces annotation effort by a factor $\log_2 n$, n being the number of alternatives to rank. In the specific case, for $n = 6$ the number

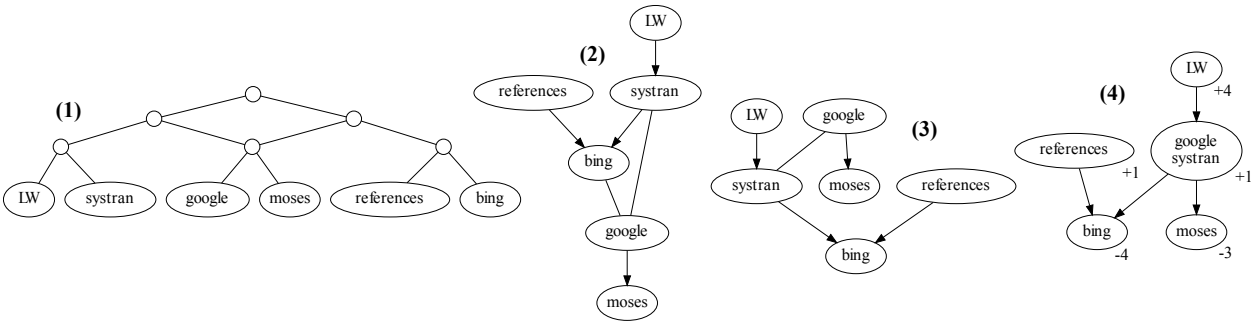

Figure 1: Tournament-based translation ranking. By annotating pairwise comparisons (1) an annotator builds a connected graph of the alternative translations (2). The dominance of the vertices is then employed to remove cycles from the graphs (3) and to obtain a full ranking for the alternative translations (4).

of required annotations is decreased by more than 60%. To further streamline the process, we assume any two translations which differ only in the casing to be equivalent. As a result, we can reduce the problem of fully ranking the alternative translations for the 1,882 source sentences to the explicit annotation of 10,203 relatively simple pairwise comparisons.

3.2 From ranked translations to absolute assessments

At the end of the pairwise annotation stage we have $1,882 \times 6 = 11,292$ relative assessments that we want to convert into absolute quality assessments in the form good/not-good (more specifically: adequate/non-adequate).

To reduce annotation effort, we again exploit the dominance graph and set up a greedy annotation framework following the algorithm:

1. Create a list $\mathcal{D}$ in which the translation hypotheses h_i are sorted by the number of dominated nodes $\delta(h_i)$, most dominant nodes first;

2. For each $h_i \in \mathcal{D}$, if h_i has not been labeled yet:

   (a) Ask the annotator if h_i is a good translation for the source sentence s, and label h_i accordingly

   (b) If h_i is labeled as inadequate, then for each $h_j \in \mathcal{D}$ so that h_j is dominated by h_i, automatically label h_j as inadequate.

That is, we assume that if h_i dominates h_j according to the pairwise comparisons and h_i is inadequate, then also h_j must be inadequate. With respect to the example in Figure 1(4), the most dominant node is the translation generated by LW, having $\delta(h_{LW}) = \hat{f}(h_{LW}) = 4$. In this case, if the user annotates LW's translation as inadequate we can assume that all the dominated translations (i.e., Google, Systran, Moses and Bing) are inadequate as well. The only translation for which we would not be able to take an automatic decision is the reference translation, since it is not dominated by LW, and therefore we would also ask the annotator to evaluate this node. On the contrary, if the annotator decides that LW is an adequate translation we cannot automatically infer the adequacy for the dominated translations: in fact, we know that they are not *as good as* LW, but they could be acceptable nonetheless, so more annotations would be needed.

There is another step that we take to further reduce the effort, and it involves exploiting the fact that we included reference translations in the mix. In fact, we can assume that 1) reference translations are adequate, and that 2) all the translations that dominate (or are in a tie with) a reference translation are at least as good as the reference. Therefore, for every graph we can automatically mark as adequate the reference translation and all the vertices dominating it. In the most conservative scenario, i.e., assuming that in a graph there are no ties (so we have 6 vertices) and that no translation ever dominates the reference node, this simple strategy already reduces annotation effort by 1/6. As we will discuss in section 4, in the real case the effort reduction is even higher. As an example, for the graph in Figure 1(4) the effort reduction would be 1/5 (i.e., 20%) even though the reference translation is not dominated by any other vertex, since two translations are equiva-

lent (Google and Systran) and therefore only 5 vertices have to be annotated.

4 Evaluation

In this section we provide empirical evidence in support of the graph-based annotation methodology. We will start by documenting the inter-annotator agreement and the effort reduction in the two steps of the annotation, respectively in sections 4.1 and 4.2; then, in Section 4.3 we will show how a very simple, heuristic harmonization process can give us very positive clues about the consistency between the two stages of the annotation process; in Section 4.4 we will compare the graphical annotation vs. a traditional, full-ranking methodology to demonstrate how the former can be more efficient and ensure higher agreement among the annotators; Finally, in Section 4.5 we will carry out a detailed error analysis to understand how the method could further be improved.

4.1 Pairwise ranking

This annotation was carried out by 16 native Spanish speakers. We set apart 10 source sentences (60 triplets) to measure the inter-annotator agreement. These sentences have been cherry-picked so as to constitute a varied and representative set of the available data, and include especially difficult, technical and noisy translation requests. These sentences have been annotated by all the 16 annotators. All the other sentences have been annotated by only one annotator. To increase consistency within the annotation, all the triplets relative to the same source sentence are assigned to the same annotator.

The absolute majority of the annotators agreed on the ternary decision in 80% of the cases (MC class ≥ 9), and in 54.24% of the cases at least 2 out of 3 annotators took the same decision (MC class ≥ 11). Consider that the probability of obtaining a random MC class ≥ 9 is 15%, which is considerably lower than the observed value. Only in 20.34% of the cases the most popular option was not selected by the absolute majority of annotators (i.e., 9). Cohen's κ, measured between the two most prolific annotators, is κ=0.55. These figures show that, even though the task is a difficult one, the problem definition is precise enough to allow for a good inter-annotator agreement. Especially in the light of the heterogeneity of the inter-annotator set, these results, which can be regarded as a lower bound of the actual inter-annotator agreement, confirm the accuracy of the annotation process.

System	Average Dominance
References	1.26
Google	0.27
Bing	0.14
Languageweaver	-0.46
Systran	-0.55
Moses	-0.67

Table 1: Average dominance of the 6 translation sources.

In Section 2 we pointed out how the frequency of cycles in the graphs can be used as an indicator of the consistency of the annotation. In this respect, we have observed cycles in only 9.72% of the graphs, with a consistency measure (Formiga et al., 2010) (based on a cycle density index) of 96.17%, stating that this percentage of translation candidates are consistent and do not cause cycles in the graphs. These figures strongly suggest that the annotation methodology enforces high self-consistency.

Another useful indicator of annotation quality is the ranking of the 5 translation systems (plus the references) that we can obtain by averaging the dominance of translations across all the 1,882 sentences, as shown in Table 1. Reference translations have by far the highest dominance, meaning that they have a tendency to be ranked quite high; Google and Bing, having access to huge amounts of web-log data for training and tuning, perform quite well on this dataset; at the bottom of the list we find the Moses baseline system, which being trained on well-formed, generally long and domain-specific (Europarl, UN and newswire) sentences does not perform adequately on the unpredictable, conversational texts in the dataset. These results are consistent with our expectations and are another positive clue of the quality of the partial ordering approach.

4.2 Absolute quality assessments

This activity was carried out by two native Spanish speakers with good command of English. Inter-annotator agreement was calculated on a shared set

Absolute quality assessments	Num	Red(%)
Translations in the dataset	11,292	-
Collapsed due to identity/ties	4,622	40.93
Unique vertices	6,670	-
Automatic assessments, of which	3,040	45.58
Reference domination	2,372	35.56
Inadequacy propagation	668	10.01
Manual annotations done	3,630	67.85

Table 2: Effort reduction observed on the absolute assessment annotation task thanks to the graphical approach.

of 66 translations obtained from 15 randomly selected source sentences. Cohen's κ between the two annotators is 0.56, Pearson's correlation is 0.61 and Spearman's correlation is 0.45. Note that Kappa is computed from the ternary decisions (A>B, A<B, A=B), while the Pearson's and Spearman's correlations are computed from the final ranks.

All these indicators show a substantial agreement between the annotators on a difficult task performed on a very eclectic dataset.

In section 3.2 we discussed the strategies that we employed to reduce annotation effort during this stage. Table 2 shows the empirical effort reduction that we observed. After building the ranked graphs, we are able to collapse 4,662 translation vertices due to identity (i.e., same surface form) or ties between translations. This first step already reduces annotation effort (in terms of sentences to be annotated) by almost 41%. Of the 6,670 vertices left, more than 45% can be annotated automatically: in 2,372 cases, the translations can automatically be labeled as adequate either because they are reference translations (there are 1,882 reference translations in the dataset, one per source sentence) or because they dominate/are in tie with the reference vertex; 668 more vertices can be automatically labeled as inadequate due to the greedy propagation of the inadequacy assessments, as discussed in Section 3.2. In the end, the annotation of the whole dataset required only 3,630 manual annotations, with an actual effort reduction close to 68%.

4.3 Harmonization

The directed graphs that we use as the basis for the absolute adequacy annotation are not completely connected, as in the example in Figure 1(4). As a consequence, it may happen that a higher-ranked translation is marked as inadequate, whereas a lower ranked translation is marked as adequate. For example, in the case of Figure 1(4) the annotator may label the translation provided by Moses (ranked 3rd) as inadequate, and the translation by Bing (ranked 4th) as adequate.

To overcome such apparent inconsistencies, we perform a post-processing step in which we *harmonize* the ranks by re-sorting the purely dominance-based ranks so that all the adequate translations are never ranked lower than an inadequate translation. After re-arranging the translations, we simply update their ranks so that:

- the best translation is ranked 1 and the ranks grow monotonically;
- two adjacent translations either have the same rank, or the difference between their ranks is 1.

For example, consider the five translations and their original ranks (in parentheses) $[a^{(1)}, b^{(2)}, c^{(3)}, d^{(3)}, e^{(4)}]$, of which only two, b and c, are marked as adequate. After moving the two adequate translations at the beginning of the list, i.e., $[b^{(2)}, c^{(3)}, a^{(1)}, d^{(3)}, e^{(4)}]$, we update the ranks to comply with the conditions above, and obtain $[b^{(1)}, c^{(2)}, a^{(2)}, d^{(3)}, e^{(4)}]$.

More formally, let $\mathcal{D}$ be the list of translations for a source sentence s, indexed from 0. Let r_i be the rank of $\mathcal{D}[i]$ according to its dominance, and $q_i \in \{0, 1\}$ be its absolute quality assessment (0 for adequate, 1 for inadequate). The harmonized ranks r'_i are obtained with the following algorithm:

1. Sort $\mathcal{D}$ by r_i, then again by q_i with a stable algorithm;
2. Initialize $r'_0 \leftarrow 1$, $r'_i \leftarrow r_i \ \forall i \in [1, n)$;
3. for $i \in [1, n)$, do:
   (a) if $r'_i < r'_{i-1}$, then $r'_i \leftarrow r'_{i-1}$;
   (b) else, while $r'_i > r'_{i-1} + 1$:
      i. for $j \in [i, n)$, do $r'_j \leftarrow r'_j - 1$

Method	Avg	Dev
Spearman correlation	0.98	0.06
Mean Absolute Error (MAE)	0.05	0.17
Root Mean Squared Error (RMSE)	0.08	0.23

Table 3: Agreement between dominance-based and harmonized ranks.

This post-processing step allows us to estimate the agreement between the two annotation stages, which we can measure by means of the correlation between the original and the harmonized ranks. To this end, we average the Spearman correlation, the Mean Absolute Error and the Root Mean Squared Error between the two set of ranks for all the sentence. The figures that we obtain, listed in Table 3, show very high correlation and negligible differences in the ranks, and demonstrate a considerably high consistency among the two different stages of the annotation process.

4.4 Full-rank vs. Pairwise ranking

To evaluate the graphical strategy against the traditional, full-ranking approach we randomly selected 20 source fragments (120 translations) and divided them in two groups of 10 translations each, A and B. The fragments have been selected among those of length between 10 and 60 words so as to be balanced with respect to the length (i.e., there are 4 sentences with length between 10 and 20, 4 between 20 and 30 and so on). Two annotators have been asked to rank the translations in group A by using pairwise comparisons, and to rank those in group B by means of full rankings. For full-ranking annotations, the annotators are only required to explicitly rank all the translations, ties being allowed. In particular, they are not forced to use a specific sorting algorithm and they are free to devise their own strategy to annotate the data as accurately and efficiently as possible. After one month[5], the annotators re-annotated the same sentences, this time by switching method, i.e. they used full-rankings for B and pairwise comparisons for B. Combining these annotations we can estimate: 1) the consistency of the ranks obtained with the two-methods, i.e., to what extent the two methods produce comparable ranks; and 2) the agreement between the two annotators when using each of the two methods.

[5]We wanted to put as much time as possible between the two iterations in order to reduce the risk that during the second iteration the annotators would remember the decisions taken during the first one.

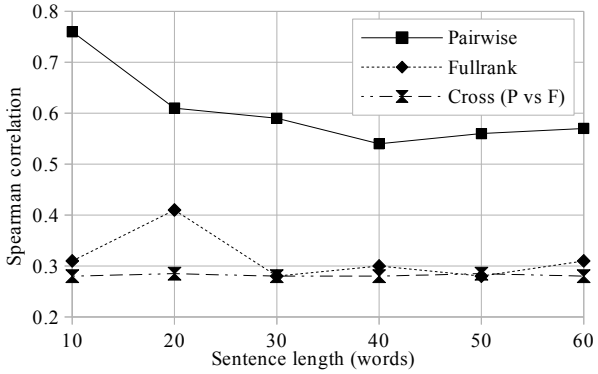

Figure 2: Inter-annotator and cross-method agreement with respect to sentence length.

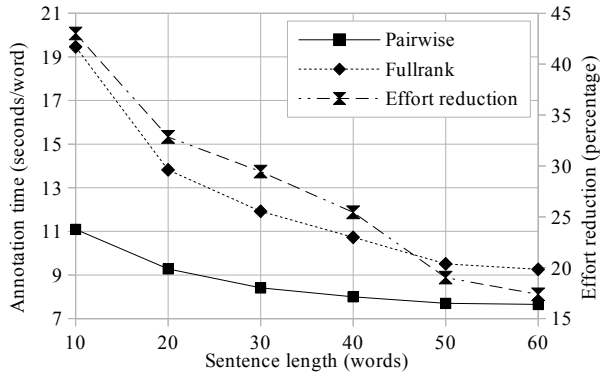

Figure 3: Per-source-word annotation time and effort reduction with respect to sentence length. X-axis depicts a cumulative value (e.g. 60 means "60 words or less").

4.4.1 Accuracy and consistency

Figure 2 shows the inter-annotator agreement (measured in terms of Spearman correlation) between the two annotators when using the full-rank (diamond) or the pairwise (square) comparisons. The agreement is shown with respect to sentences of different maximum length, e.g., a data point with $x = 40$ describes all the sentences having length $l \leq 40$. The chart shows how pairwise comparisons result in much more consistent annotations, with Spearman correlation ranging between 0.5 and 0.6 almost independently of the length of the sentences. The third series in the chart (marked by two

triangles) shows the consistency of the two annotators across the two different methods. This figure is constantly low ($\sim$0.28), telling that the two methods are producing quite different rankings in the end. This phenomenon has two possible explanations:

1. The explicit annotation of the full ranking is a complex task for non-trivial cases. The annotators get frustrated and, as a result, tend to relax the annotation criterion and do their job less carefully. We call this the *fatigue* hypothesis;

2. The small number of comparisons carried out in the pairwise approach is actually insufficient to correctly establish the full ranking of the translations. Not being completely connected, the graphs are missing some comparisons that would be necessary to establish the correct ranking of the alternatives. We call this second hypothesis the *topological* hypothesis.

The consistently higher agreement of the pairwise approach is already a good indicator of the validity of the fatigue hypothesis, but is not sufficient to completely rule out the topological hypothesis. In order to do so, we would need to show that, assuming a perfect annotator taking the same decisions regardless of the method employed, the pairwise method would produce the same results as the full-ranking strategy.

To this end, we ran a completely unsupervised experiment based on the human quality assessments released as part of the WMT 2010 evaluation campaign (Callison-Burch et al., 2010). For each set of ranked translations, we created the minimum number of pairs required to build a connected graph of the alternatives, and used the ranks to simulate a pairwise decision about each selected translation pair: if the two translation have the same rank, we annotate the pair as a tie, otherwise we mark one as better than the other depending on their respective ranking. Then, we process the graph as already described in Section 3 and obtain a complete ranking of the translations only based on the pairwise comparison. We considered all the 860 source sentences for which a judge ranked five alternative translations, and measured the Pearson's correlation between the original and the reconstructed ranks. We observed an average correlation of $r = 0.93$ and a standard deviation of 0.08, meaning that the reconstructed ranks are for practical purposes identical to the original ones and demonstrating that the topological hypothesis can be ruled out as the cause of the rank disagreement previously observed. The relative rankings established without a direct comparison (i.e., those solely inferred from dominance differences) are correct in 85.4% of the cases. Even more, the higher inter annotator agreement that we observed with pairwise comparisons makes us believe that the resource would actually be more accurate and more consistently annotated if it had been produced by means of simple pairwise comparisons.

4.4.2 Efficiency

Figure 3 shows, on the left axis, the averaged per-source-word annotation time (i.e., the time required to annotate a whole set of alternative translations, divided by the number of words in the source sentence) with respect to sentences of different length when using the two annotation strategies. We can observe how pairwise ranking (squares) is constantly faster than full ranking (diamonds), even though the difference between the two methods decreases with the length of the sentences considered. On the right axis we also plot the effort reduction (two triangles) that can be achieved by using the pairwise comparisons. Effort reduction is higher than 30% for translations of 30 words of less, with an improvement of almost 45% for very short texts ($\leq$ 10 words). Globally (length $\leq$ 60), the effort reduction in terms of per-source-word annotation time has been measured at 17.38%. In this effort reduction results we assume that there is a fixed time requirement to take a decision which is not proportional to the length of the sentence (initial offset).

This set of annotations has been done with pen and paper, meaning that the annotators could underline and mark parts of the texts and approach the annotation in a more structured way. The annotators have commented that they perceived the pairwise comparison as extremely streamlined when annotating medium-short sentences, whereas they found it more cumbersome when the source sentence is long. In fact, in the pairwise case the annotators had to read 12 translations (6 pairs), while in the full-ranking approach they only had to read 6. In this respect, the pen-and-paper setting is much friendlier

to the full-ranking method, also because the ability to mark some parts of the longer sentences allowed the annotators to re-use the same information across several comparisons. Furthermore, annotators commented that for longer sentences, lexical aspects might become less relevant because more evident syntactic differences (or errors) emerge.

4.5 Error analysis

With respect to a full-ranking strategy, the pairwise ranking approach suffers from the fact that in order to reduce the number of triplets to be annotated we are only able to induce a partial ordering among the translations. This adds a potential source of imprecision to the typical problems of annotation such as human errors, inconsistencies or the change of criterion during the course of the annotation. In Section 4.4.1 we have already demonstrated how the effect of this topological bias is indeed very marginal. In this section, we select some practical example to analyze this aspect in more detail.

The topological problem usually arises when two proper solutions (competing ones) are not directly compared. In this case, one solution may obtain better rank despite it would be annotated as worse in a direct comparison. An example can be found in Figure 1(4): the reference translation, possibly the best in the lot, has been involved in only one comparison, which happened to be against the worst translation. This lack of comparisons assigns the reference a dominance value of 1, putting it in the same rank as Google and Systran and one place below LW (with dominance 2). If the reference translation had been compared, say, against LW, then the final ranking might have been very different. In the Genetic Algorithms framework, this problem is generally overcome by subsequent iterations in which the annotators are explicitly required to compare competing solutions, at the cost of higher annotation effort.

The dominance-based and the harmonized ranks (see Section 4.3) differ in 212 cases out of 1,882, i.e. 11.26%. Two expert annotators analyzed 50 such cases individually, in order to isolate the specific factors that can hinder the reliability of the graphical ranking scheme. We found that the topological problem to be present in 60% of the cases, most of which (43% of all conflicts) could be resolved by explicitly annotating the direct comparison between two competitive alternatives. The remaining 40% of conflicts can be ascribed to inconsistent human annotations, either during the first or the second stage of the annotation. Most human inconsistencies (24% of all conflicts) are due to a change of criterion, while the other inconsistencies (16%) are caused by human errors that we can attribute to fatigue, lack of concentration or other psychological or environmental factors. The harmonized ranking was able to completely fix 20% of all conflicts between the pairwise rankings, and to partially correct 18% of the conflicts. In this respect, the absolute annotation somehow supplements for the lack of a second run of pairwise annotations.

5 Conclusions

In this work we have presented and analyzed a graph-based annotation scheme employed to annotate a real corpus of English–Spanish automatic translations with relative (ranking) and absolute (adequate/non-adequate) translation quality assessments. The annotation methodology has been borrowed from the context of Active Interactive Genetic Algorithms (aiGAs) and has been applied, for the first time, to the creation of annotated resources for MT. The method is based on decomposing the full ranking problem into a small set of pairwise comparisons which lead to a connected graph among translation alternatives. Afterwards, the dominance relation on the constructed graphs can be easily used to automatically derive complete rankings of the translation alternatives. By thoroughly comparing the graph-based approach to the direct annotation of complete ranks among translation candidates, we showed that the new methodology is able to produce better ranks (in terms of higher inter-annotator agreement) at a reduced time cost (especially for short-medium length sentences). Apart from these two good properties, we also presented a pipelined application of the graph-based method with a procedure for producing absolute quality assessments. The combined application provided a 67% reduction of the annotation effort in the latter stage, and helps fixing topological issues that arise in some specific cases. According to previous aiGA findings, that we confirmed with an empirical error analysis, these special cases can generally be

fixed by means of one more comparison among two isolated top-ranked alternatives. Last but not least, we have demonstrated how the proposed framework would have made it possible to produce all the WMT 2010 human quality assessments with considerably reduced effort without sacrificing the accuracy of the result.

6 Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This research has been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 247762 (FAUST project, FP7-ICT-2009-4-247762) and by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03).

References

Francesc Alías, Lluís Formiga, and Xavier Llorà. 2011. Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept. *Speech Communication*, 53(5).

C.O. Alm and X. Llorà. 2006. Evolving Emotional Prosody. In *Proc. of ICSLP'06*, pages 1826–1829, Pittsburgh, PA (USA).

C.O. Alm. 2011. Subjective natural language problems: motivations, applications, characterizations, and implications. In *Proceedings of ACL-HLT'11*, pages 107–112.

Rafael E. Banchs and Haizhou Li. 2011. AM-FM: A Semantic Framework for Translation Quality Assessment. In *Proceedings of ACL'11*, pages 153–158, Portland, Oregon, USA, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan, editors. 2010. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden.

C.A. Coello. 2000. An updated survey of ga-based multiobjective optimization techniques. *ACM Computing Surveys (CSUR)*, 32(2):109–143.

K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. 2000. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. *Lecture notes in computer science*, 1917:849–858.

L. Formiga, F. Alías, and X. Llorà. 2010. Evolutionary Process Indicators for Active IGAs Applied to Weight Tuning in Unit Selection TTS Synthesis. In *Proceedings of the IEEE CEC'10*, Barcelona.

Jesús Giménez and Lluís Màrquez. 2010. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24:209–240. 10.1007/s10590-011-9088-7.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: MT Summit'05*, pages 79–86, Phuket, Thailand.

X. Llorà, F. Alías, L. Formiga, K. Sastry, and D.E. Goldberg. 2005a. Evaluation consistency in iGAs: User contradictions as cycles in partial-ordering graphs. *Urbana*, 51:61801.

Xavier Llorà, Kumara Sastry, David E. Goldberg, Abhimanyu Gupta, and Lalitha Lakshmi. 2005b. Combating User Fatigue in iGAs: Partial Ordering, Support Vector Machines, and Synthetic Fitness. In *Proceedings of GECCO'05*, pages 1363–1370, New York, NY, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL'02*, pages 311–318, Stroudsburg, PA, USA. ACL.

Vilfredo Pareto. 1896. *Cours d'economie politique*, volume I, II. F. Rouge.

Daniele Pighin, Lluís Màrquez, and Lluís Formiga. 2012. The FAUST Corpus of Adequacy Assessments for Real-World Machine Translation Output. In *Proceedings of LREC'12*, Istanbul, Turkey, may. ELRA.

Alexandre Rafalovitch and Robert Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *Proceedings of the MT Summit XII*, pages 292–299. International Association of Machine Translation, August.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the Confidence of Machine Translation Quality Estimates. In *Proceedings of Machine Translation Summit XII*, Ottawa, Canada.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of LREC'10*, Valletta, Malta.

H. Takagi. 2001. Interactive Evolutionary Computation: Fusion of the Capabilities of the EC Optimization and Human Evaluation. *Proceedings of the IEEE*, 89(9):1275–1296.

Yi-Hsuan Yang and Homer H. Chen. 2009. IMR: Interactive Music Recommendation via active interactive Genetic Algorithm. In *International Workshop on Computer Music and Audio Technology*.

Annie Zaenen. 2006. Mark-up barking up the wrong tree. *Comput. Linguist.*, 32(4):577–580, December.