Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification

Raphael Rubino
Universität des Saarlandes
Saarbrücken, Germany

Ekaterina Lapshinova-Koltunski
Universität des Saarlandes
Saarbrücken, Germany

Josef van Genabith
Universität des Saarlandes
DFKI
Saarbrücken, Germany

e.lapshinova@mx.uni-saarland.de
{raphael.rubino,josef.vangenabith}@uni-saarland.de

Abstract

This paper introduces information density and machine translation quality estimation inspired features to automatically detect and classify human translated texts. We investigate two settings: discriminating between translations and comparable originally authored texts, and distinguishing two levels of translation professionalism. Our framework is based on delexicalised sentence-level dense feature vector representations combined with a supervised machine learning approach. The results show state-of-the-art performance for mixed-domain translationese detection with information density and quality estimation based features, while results on translation expertise classification are mixed.

1 Introduction

Translations, regardless of the method they were produced with, are different from their source texts and from originally authored comparable texts in the target language. This has been confirmed by many linguistic studies on translation properties commonly called *translationese* (Gellerstam, 1986). These studies show that translations tend to share a set of lexical, syntactic and/or textual features distinguishing them from non-translated texts. As most of these features can be measured quantitatively, we are able to automatically distinguish translations from originals (Baroni and Bernardini, 2006; Ozdowska and Way, 2009; Kurokawa et al., 2009). This is useful for Statistical Machine Translation (SMT), as language and translation models can be improved if the translation direction and status of the data (translation or original) is known (Lembersky, 2013).

Research on translationese has recently focused on exploring features capturing aspects of translationese such as simplification, explicitation, convergence, normalisation and shining-through (Volansky, 2012; Ilisei, 2012). Here we extend this work as follows: (i) we investigate the impact of information density and surprisal features, (ii) we explore the use of features used in machine translation quality estimation (Blatz et al., 2003; Specia et al., 2010), (iii) we explore classification between originally authored text and trainee and professional translation, as well as between professional and trainee translation. In order to avoid biasing classification by topic content, throughout our experiments we use fully delexicalised features, resulting in dense vector representations (rather than sparse vectors, where the size of the vectors can be up to and in fact exceed the size of the vocabulary). We show that information theory as well as translation quality estimation inspired features achieve state-of-the-art results in mixed-domain original vs. human translation classification.

Languages provide speakers with a large number of possibilities of how they may encode messages. These include the choice of phonemes, words, syntactic structures, as well as arranging sentences in discourse. Speakers' decisions regarding these choices are influenced by diverse factors: cognitive processing limitations can impact variation in linguistic encoding across all linguistic levels. Text production conditions, including monolingual vs. multilingual settings, can influence this variation: in

Proceedings of NAACL-HLT 2016, pages 960–970,
San Diego, California, June 12-17, 2016.

translation, choices can be shaped by aspects of both the source and the target language.

Contrastive studies have shown that information density is distributed differently in English and German (Doherty, 2006; Fabricius-Hansen, 1996). These contrasts may impact translation, and in case of source language shining through[1], we would expect to observe differences between translations and comparable originals in terms of information density. Additionally, translations are often more specialised and more conventionalised than originals (excluding translation of fictional texts). In this paper we investigate whether and to what extent information density based features are useful in human translation classification.

Quality estimation (QE) (Blatz et al., 2004; Ueffing and Ney, 2005) is the attempt to learn models that predict machine translation quality without access to a reference translation at prediction time. Translation, manual or automatic, is always a process of transforming a source into a target text. This process is prone to error. In this paper we explore whether and to what extent the extensive research on QE can be brought to bear on the problem of human translation vs. originals classification, and in particular the discrimination between novice and professional translation output.

Below we explore the ability of our features to distinguish between 1) non-translated texts and translations by professionals, 2) non-translated texts and translations by translator trainees, and 3) the two translation varieties that diverge in the degree of translation experience. We report results in terms of accuracy and f-score, and provide a feature analysis in order to understand the role of the information density and QE inspired features in the task.

The paper is organised as follows: related work is presented in Section 2. The experimental setup is detailed in Section 3, followed by the results and analysis in Section 4. A discussion about our results compared to previous work is given in Section 5. Finally, conclusion and future work are provided in Section 6.

[1]If translations demonstrate features more typical for the source language, see e.g. Teich (2003).

2 Related Work

We briefly review previous work on translationese, information density, machine translation quality estimation and studies on human translation expertise.

2.1 Translationese

A number of corpus-based studies on translation have shown that it is possible to automatically predict whether a text is an original or a translation (Baroni and Bernardini, 2006; Koppel and Ordan, 2011). These approaches are based on the concept of translationese – a term coined to capture the specific language of translations by Gellerstam (1986). The idea is that translations exhibit properties which distinguish them from original texts, both the source texts of the translation and comparable texts originally authored in the target language. Baker (1993; 1995) claimed these properties to be universal, i.e. (source) language-independent, emphasising general effects of the process of translation.

However, translationese includes features involving both source and target language. Most linguistic studies distinguish *explicitation* – a tendency to spell things out rather than leave them implicit and *implicitation* (the opposite effect), *simplification* – a tendency to simplify the language used in translation, *normalisation* – a tendency to exaggerate features of the target language and to conform to its typical patterns, *levelling out* or *convergence* – a relatively higher level of homogeneity of translated texts compared to non-translated ones, and *interference* or *shining through* (e.g. Teich (2003)). While simple lexicalised features including word tokens and character n-grams can produce near perfect classification results for in-domain data (Avner et al., 2014), a significant amount of work has gone into devising features that can capture presumed linguistic aspects of translationese (Volansky, 2012). Rabinovich et al. (2015) explore unsupervised discrimination of translations based on principal components analysis for dimensionality reduction followed by a clustering step. The method is robust to unbalanced and heterogeneous datasets, which may be useful to handle mixed domain, genre and source of data, a common situation when training language and translation models.

Automatic classification of original vs. translated

texts has applications in machine translation, especially in studies showing the impact of the nature (original vs. translation) of the text in translation and language models used in SMT. Kurokawa et al. (2009) show that taking directionality into account when training an English-to-French phrase-based SMT system leads to improved translation performance. Ozdowska & Way (2009) analyse the same language pair and demonstrate that the nature of the original source language has an impact on the quality of SMT output. Lembersky et al. (2012) show that BLEU scores can be improved by language models compiled from translated texts and not from comparable originally authored ones.

2.2 Information Density

In a natural communication situation, speakers tend to exploit variations in their linguistic encoding – modulating the order, density and specificity of their expressions to avoid informational peaks and troughs that may result in inefficient communication. This is often referred to as the *uniform information density* hypothesis (Frank and Jaeger, 2008). The information conveyed by an expression can be quantified by its *surprisal*, a measure of how predictable an expression is given its context. Simplification and explicitation may impact the average information density measured on translated texts compared to comparable originally authored ones in the same language. Source language interference should result in peaks of measured surprisal values in translated texts, while the information density may remain uniform in originals.

According to Hale (2001), a surprisal model allows the estimation of the probability of a parse tree given a sentence prefix. Levy (2008) showed that a lexical-based surprisal measure can be obtained by computing the negative log probability of a word given its preceding context: $S = -\log P(w_{k+1}|w_1 \dots w_k)$. Following Demberg et al. (2013), we estimate surprisal in three ways, at the word, part-of-speech and syntax levels, based on n-gram language models and language models trained on unlexicalised part-of-speech sequences and flattened syntactic trees. Note that all resulting feature vectors do not represent lexical information but information theoretic surprisal measures.

2.3 Quality Estimation

Machine translation QE is the process of estimating how accurate an automatic translation is through characteristic features of the source and target texts, and (possibly) also the translation engine, with a supervised machine learning setting to estimate quality scores. QE can be applied at the word, sentence and document level (Gandrabur and Foster, 2003; Ueffing et al., 2003; Blatz et al., 2003; Scarton and Specia, 2014).

Many different delexicalised dense features have been explored in previous work on QE, including language and topic models, n-best lists, etc. (Quirk, 2004; Ueffing and Ney, 2004; Specia and Gimenez, 2010; Rubino et al., 2013a). It has been shown that the performance of a supervised classifier to distinguish between originals and automatic translations is correlated with the quality of the machine translated texts (Aharoni et al., 2014): low quality translation, containing grammatical and syntactic errors, as well as incorrect lexical choices, are robust indicators of automatic translations. In the case of human translation, to the best of our knowledge, there are no empirical studies on the level of professional expertise in the translation process and its correlation with the performance of a translationese classifier.

2.4 Translator Experience

Jääskeläinen (1997) describes translational behaviour of professionals and non-professionals who perform translation from English into Finnish. Carl and Buch-Kromann (2010) apply psycholinguistic methods in their analysis. They present a study of translation phases and processes for student and professional translators, relating translators' eye movements and keystrokes to the quality of the translations produced. They show that the translation behaviour of novice and professional translators differs with respect to how they use the translation phases. Englund Dimitrova (2005) develops a combined process and product analysis and compares translators with different levels of translation experience, but concentrates only on cohesive explicitness.

Most of these works are rather process-oriented than product-oriented, which means that features of translated texts are rarely taken into account. How-

ever, some of the findings are valuable for the analysis of translated texts. For instance, Göpferich & Jääskeläinen (2009) find that with increasing translation competence, translators focus on larger translation units, which can impact the choice of linguistic encoding translators use.

3 Experimental Setup

Our experiments are designed to investigate underexplored topics focusing on (i) information theoretic and (ii) machine translation QE features in translation classification. We use dense vector representations with fully delexicalised features and investigate three hypotheses:

1. originals & professional translations should be close in terms of quality and thus more difficult to separate automatically,
2. originals & student translations should be distant in terms of quality and thus easier to classify,
3. professional & student translations should both contain translationese features and thus may be very difficult to differentiate.

3.1 Supervised Classification

In order to train a classifier and predict binary labels on unseen data, we use a dense vector sentence-level representation associated with a class $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, l$ (l is the number of training instances) with $\mathbf{x}_i \in R^n$ (n is the size of a dense vector) and $y \in \{-1, 1\}^l$. We train classification models with a support vector machine SVM (the C-SVC implementation in LIBSVM (Chang and Lin, 2011)) as a quadratic optimization problem:

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \tfrac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i\ , \\ \text{subject to} \quad & y_i(\omega^T\phi(x_i) + b) \geq 1 - \xi_i\ , \quad \xi_i \geq 0\ . \end{aligned}$$

ϕ is a kernel function and allows the projection of training data to a higher dimensional space. We use the radial basis function (RBF) kernel, as it produced the best empirical results compared to linear and polynomial kernels. We predict the class for unseen instances x as follows:

$$f(x) = \text{sgn}(\omega^T\phi(x) + b).$$

Corpus	Token (M)	Sentence (k)
Europarl Originals	4.1	155.5
Literature Originals	1.3	48.1
Literature Translations	1.4	45.8
Politics Originals	0.2	9.7
Politics Translations	0.2	8.7

Table 1: Details of the corpora used to train language and n-gram frequency models for originally authored texts and translations.

Two hyper-parameters have to be set for C-SVC with the RBF kernel: the regularisation parameter (or penalty) C and the kernel parameter γ. We use grid-search to find optimal values, performing a 5-fold cross-validation on the training data. To avoid over-fitting, we use a held-out development set to evaluate the models obtained.

3.2 Datasets

The datasets used in our experiments are separated into two subsets: corpora used to extract features and corpora used to train, tune and test our classifiers. The former are taken from the publicly available bilingual English-German parallel corpora consisting of parliamentary proceedings, literary works and political commentary, compiled by (Rabinovich et al., 2015). These corpora are used individually to train language models and compute n-gram frequency distributions. Basic corpus statistics are presented in Table 1. The latter ones are composed of German texts, taken from the VARTRA corpora (Lapshinova-Koltunski, 2013), which were either originally written in German (originals) or translated from English (translations).

Originals and translations belong to the same genres and registers and can be considered comparable. They include a mixture of literary, tourism and popular-scientific texts, instruction manuals, commercial letters and political essays and speeches. The VARTRA translations are split in two sets: one produced by professional translators, and one produced by translator trainees. Details are presented in Table 2. We extract balanced subsets of training, tuning and testing data containing three, one and two thousands sentences, respectively, of each type.

Corpus	Token (k)	Sentence (k)
Originals	121.7	6.0
Professional Translations	125.2	6.0
Student Translations	126.2	6.0

Table 2: Details of the comparable corpora used as training, development and test sets for the originals versus translation classification experiments.

3.3 Feature Sets

For classification, input text is represented as a set of feature vectors. The features capture aspects of information density and translation QE. Throughout we use unlexicalised lower-dimensional dense vectors rather than high-dimensional lexicalised sparse vectors to minimize the input of specific content on classification results. We extract a total of 778 features[2] and separate them into four subsets corresponding to broad but distinct characteristics of original and translated sentences: surface and distortion features are related to QE, surprisal and complexity features are inspired by information theory.

Surface Features - 13 surface features based on meta representations of sentences' lexical form. Features include sentence and average word length, the number of word tokens and number of punctuation marks. Three case-based features capture the number of upper-cased letters and words, and a binary feature indicates whether a sentence starts with an upper-case character. Another binary value encodes whether the sentence ends with a period. Two features are obtained from the ratio between the number of upper-cased and lower-cased letters, the number of punctuation marks and the length of the sentence. Finally two features capture the number of periods merged with words and words with mixed-case characters.

Surprisal Features - 225 features based on the surprisal measure presented in Section 2.2 are extracted using language models trained on words, delexicalised part-of-speech and flattened syntactic trees. The language models are trained on individual[3] corpora presented in Table 1. We extract n-gram ($n \in [1;5]$) log-probabilities and perplexities, with and without the tags indicating the beginning and ending of sentences, using the SRILM toolkit (Stolcke et al., 2011).

Complexity Features - 315 features based on n-gram frequencies, indicating how frequent the lexical choices, part-of-speech and flattened syntactic sequences present in the text to be classified are. As for the surprisal features, we use the same originally authored and translated texts individually to extract n-grams frequency quartiles. We extract the percentage of n-grams ($n \in [1;5]$) occurring in each quartile. Frequency percentages are averaged at the sentence level, leading to 4 features per sentence (one per quartile) given a value of n, for each corpus used to define the frequency quartiles. This approach allows us to avoid encoding raw n-gram features and keep a dense vector representation (Blatz et al., 2003).

Distortion Features - 225 features based on the possible distortion in lexical, part-of-speech and syntactic structures observed between originals and translations, as well as between different levels of translation experience. These features are extracted the same way as the suprisal features, but based on language models trained on sentence-level reversed text. The backward language model features are popular in translation quality estimation studies and show interesting results (Duchateau et al., 2002; Rubino et al., 2013b).

3.4 Preprocessing and Tools

All data used in our experiments are sentence-split, lower-cased and tokenised using the CoreNLP toolkit (Manning et al., 2014). The part-of-speech tags and syntactic trees required to extract some features are obtained with the same set of tools. For parsing, we use the probabilistic context-free grammar model trained on the Negra corpus (Brants et al., 2003) and described in (Rafferty and Manning, 2008), before flattening the trees as illustrated in Figure 1. Both part-of-speech and flattened syntac-

[2]Too many to list in the paper, a complete list is provided with the additional material submitted.

[3]Originally authored texts and translations are used separately in order to model their characteristics.

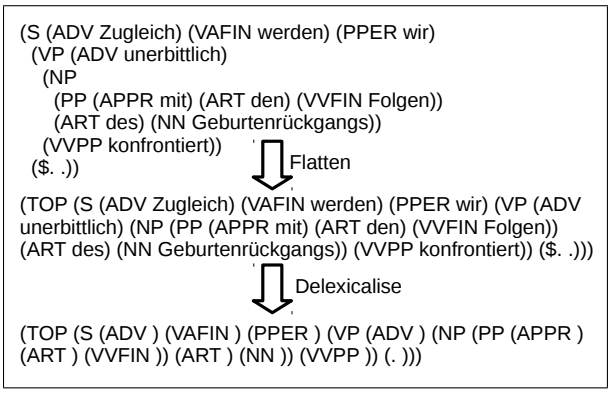


Figure 1: Flattening and delexicalising a syntactic tree.

tic trees are then delexicalised in order to remove all surface forms from the representations.

4 Results and Analysis

Below we provide details on discriminating between originally authored texts and translations, followed by the prediction of translation experience comparing professional translators and students. Finally, we evaluate feature importance with individual and ensemble feature selection techniques.

4.1 Original vs Translated Texts

Two sets of experiments are conducted to discriminate between originals and professional translations (Table 3) and originals and student translations (Table 4). For each classification task, we evaluate feature groups on the test set containing $4,000$ unseen sentences balanced over two classes, reporting overall accuracy, and also precision, recall and f-score. Finally, a classification model is trained and evaluated combining all features.

Originals vs. professional translations reaches a maximum accuracy of 70.0% using the distortion feature set with surprisal a close second at 69.2%. The difference is not statistically significant (bootstrap resampling at $p < 0.05$). They outperform the other types of features, as well as the combination of all feature types. Per class evaluation shows a similar trend with the best performing feature sets. The results show that originals and professional translations exhibit differences in terms of sequences of words, part-of-speech and syntactic tags which are captured by language model based features.

		Originals			Professional		
Feature set	Acc (%)	P	R	F	P	R	F
Surface	54.7	0.54	0.64	0.58	0.56	0.46	0.50
Surprisal	69.2*	0.66	0.77	0.71	0.73	0.61	**0.66**
Complexity	65.3	0.63	0.73	0.68	0.68	0.57	0.62
Distortion	**70.0***	0.66	0.81	**0.73**	0.75	0.59	**0.66**
All	66.5	0.64	0.74	0.69	0.70	0.59	0.64

Table 3: Accuracy, precision, recall and F-measure obtained on the originals versus professional translations classification task. Best results in bold and statistically significant winner marked with * ($p < 0.05$).

		Originals			Student		
Feature set	Acc (%)	P	R	F	P	R	F
Surface	57.8	0.58	0.58	0.58	0.58	0.58	0.58
Surprisal	69.7*	0.69	0.72	0.70	0.71	0.67	**0.69**
Complexity	65.4	0.62	0.81	0.70	0.73	0.49	0.59
Distortion	70.8*	0.69	0.75	**0.72**	0.73	0.66	**0.69**
All	**71.1***	0.69	0.76	**0.72**	0.73	0.66	**0.69**

Table 4: Accuracy, precision, recall and F-measure obtained on the originals versus student translations classification task. Best results in bold and statistically significant winner marked with * ($p < 0.05$).

The classification of originals and student translations shows that the combination of the four feature types leads to the most accurate classifier, followed by the distortion and the surprisal sets (with equivalent accuracy results at $p < 0.05$). The two latter feature sets are the best performing ones overall based on the two classification tasks. Comparing the two tasks, originally authored texts are closer to professional translations and more distant to student translations, which validates two of our hypotheses listed in Section 3.

4.2 Translation Expertise

In order to investigate whether our third assumption is correct, we perform binary classification between professional and student translations (Table 5). The results, barely above the 50% baseline, show the proximity of the two types of translations according to our feature sets, which supports our third assumption. The combination of four feature types reaches the highest accuracy, followed by the distortion and complexity sets. However, the surprisal features do not seem to be helpful in differentiating between the

professional and the student translations, compared to the two previous binary classification tasks.

This result indicates that the surprisal measure is a reliable source of information to determine whether a sentence is originally authored or a translation, but it is not reliable to separate two translations produced by translators with different levels of expertise. The features inspired by translation quality estimation do not reach high accuracy results: it seems that the difference between professional and student translations cannot be tied to properties of the surface level or lexical choices of the human translators as indirectly captured by our features.

		Professional			Students		
Feature set	Acc (%)	P	R	F	P	R	F
Surface	54.5	0.56	0.43	0.48	0.54	0.66	0.59
Surprisal	55.7	0.57	0.48	0.52	0.55	0.64	0.59
Complexity	56.0	0.56	0.55	0.56	0.56	0.57	0.56
Distortion	57.7	0.58	0.55	0.56	0.57	0.60	0.59
All	**58.7***	0.59	0.57	**0.58**	0.58	0.61	0.59

Table 5: Accuracy, precision, recall and F-measure obtained on the professional versus student translations classification task. Best results in bold and statistically significant winner marked with * ($p < 0.05$).

4.3 3-way Classification

Table 6 shows the confusion matrix obtained with the classifier trained on the combination of the four feature sets. This classifier reaches third position overall in terms of accuracy, behind the distortion and surprisal sets with first and second positions, respectively. This ranking of classifiers trained on different feature sets follows the trend observed in the originals versus professional translation binary classification task.

		Reference		
		Originals	Professional	Student
Prediction	Originals	1318	656	544
	Professional	276	699	491
	Student	406	645	965

Table 6: Confusion matrix obtained using a classifier trained on the four feature sets for the multi-class task, separating originals, professional and student translations.

The training method chosen for the multi-class problem is the *one against one*, where individual classifiers are first trained on each binary classification task before being combined to form the final multi-class classifier (Hsu and Lin, 2002). The results indicate that our feature sets distinguish originally authored texts from professional and student translations (first line of the matrix), while the professional translations are more difficult to separate from the two other types of text. Also, student translations have characteristics differing from originals and professional translations, which can be captured with our feature sets (last line of the matrix). However, the columns of the confusion matrix show that originals are not necessarily closer to professional translations, as indicated by the first column where a larger amount of gold originals are incorrectly classified as student translations. The same trend is observable in the last column. These results go against the hypothesis that originals and student translations are easier to separate, a phenomenon which does not appear for the binary classification task (originals vs. student translations).

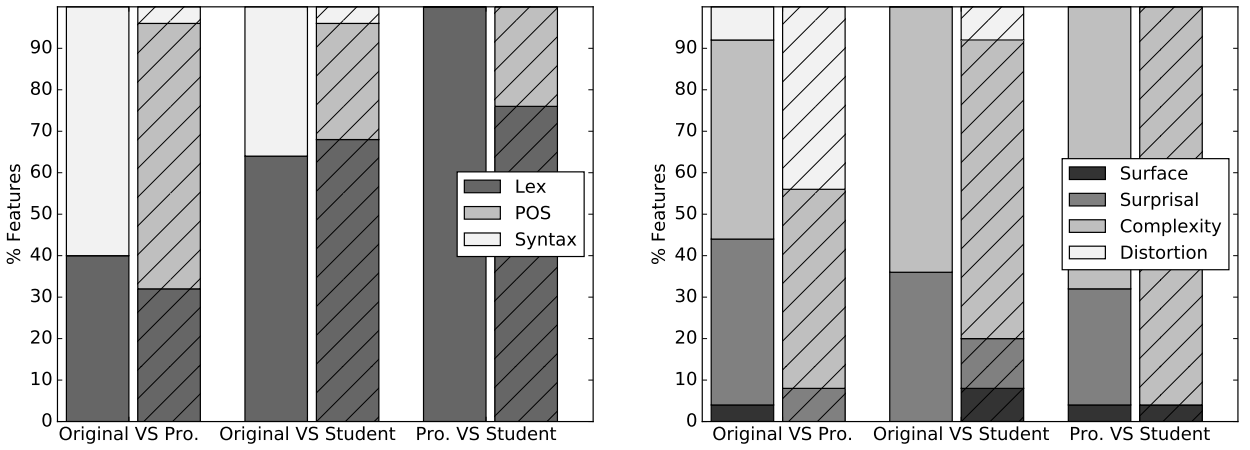
4.4 Feature Performance

Evaluating the performance of our feature sets is done by calculating the discriminative power of each feature individually which allows us to rank features according to their correlation with a class given a classification task. We follow the "f-score" measure (1) as proposed by Chen (2006):

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(\bar{x}_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(\bar{x}_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (1)$$

with training vectors x_k and $k = 1, \ldots, m$, binary classes n_+ and n_- for positive and negative instances, $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ the average of the ith feature of the whole, positive and negative instances, and $\bar{x}_{k,i}^{(+)}$ and $\bar{x}_{k,i}^{(-)}$ the ith feature of the kth positive or negative instance. The measure indicates how discriminative a feature is given a binary classification task. A drawback of the *f-score* is that it does not take into account possible feature complementarity.

We report the distribution of the top 25 features amongst the three levels of analysis: lexical, POS and syntax (Figure 2a), as well as amongst the four



(a) Lexical, POS and syntax.

(b) Surface, surprisal, complexity and distortion.

Figure 2: Distributions of the top 25 most important features according to individual discriminative power (left bars) and ensemble of randomised trees (right hatched bars).

feature types: surface, surprisal, complexity and distortion (Figure 2b). The results show that POS features are not ranked as the most discriminant ones when evaluated individually, while syntactic features are the most important ones for the originals vs. professional translation task and lexical features have the highest discriminative power for the two other tasks. When looking at the feature types, we see that complexity features, based on n-gram frequencies, are the most discriminant for the three tasks, followed by the surprisal features, while the distortion and surface features do not have a strong discriminative power. Most of the top n-gram based features rely on sequences between 1 and 3 words, indicating that higher order n-grams are not important features when considered individually. Surprisal, distortion and complexity features are based on external resources (detailed in Table 1) and the corpus of political texts translated into German is the most useful one when used to extract the complexity and surprisal features, which can be explained by the presence of political speeches and essays in the VARTRA corpus.

The results obtained on individual feature discriminative power do not reflect the ones obtained using features grouped by types. Individually, features indicating complexity based on n-gram frequencies are ranked highest. However, only a few of the distortion features appear in the discriminative ranking while this feature type reaches the highest accuracy scores on the three binary classification tasks. These results indicate that features are highly complementary within a group of a particular type, but also between different types. To capture possible relationships between features, we conduct a non-linear feature selection using the forest of randomised trees approach (Geurts et al., 2006) and present the results for the top 25 features in Figure 2 (right hatched bars).

The tree-based feature ranking method shows the complementarity of words and POS features, while the syntactic ones appear in the top 25 for the original vs. translation tasks for both levels of expertise. When looking at the feature types, the originals vs. professional task relies mainly on a mixture of distortion and complexity features, and surprisal indicators are totally absent from the top 25 for the professional vs. student task. For both tasks involving student translation, the complexity features are the most important ones, and simple surface features are useful, such as the average words occurrence per sentence or the ratio between the number of punctuation marks and the sentence length. The most useful external resource used to extract n-gram based features is again the political corpus, indicating once more the domain proximity of our datasets.

Individually, syntactic features appear to be highly discriminant when classifying between originals and translations (regardless expertise), which may indicate two translationese phenomena: simpli-

fication, translators use less complex constructions, and interference (shining through), source syntax shines through in translated texts. The ensemble ranking shows that surprisal and distortion, although not as important as complexity and distortion, are important indicators of translationese as they appear in both tasks where originals are classified against translations. These feature types are not present in the top 25 if only translated texts are classified.

5 Discussion

Previous research (Baroni and Bernardini, 2006; Volansky, 2012) has shown that high classification accuracy ($> 80\%$) can be achieved using lexicalised token n-gram sparse feature vectors. As a sanity check, we conduct a set of experiments for each of our classification tasks using token unigram frequency as features, normalised by the segment length. The vocabulary defining the feature vector dimensionality is taken from the training sets, using the data presented in Table 2 only, leading to $25,561$ features. The same classification setup as presented in Section 3 is used and we observe accuracy results reaching 78.0%, 83.3% and 65.2% for original vs. professional, originals vs. student and professional vs. student classifications respectively. For the three-way task, an accuracy score of 62.7% is reached. These results are substantially lower than the ones reported by Volansky (2012), mostly because of the text chunks size, which has a strong impact on performance as shown by Rabinovich and Wintner (2015). In our work, we classify each sentence individually as they appear naturally in the corpus, while most previous studies are based on artificial chunks of approximately $2,000$ tokens. An other explanation of the low performing unigram-based features is related to our mixed-domain setting, as it was shown that classifiers' performance drop drastically when trained on this type of features and tested on out-of-domain data (Rabinovich and Wintner, 2015).

6 Conclusion

This paper presented a first step in using information density, and especially surprisal and complexity inspired features, as well as features used in translation quality estimation, as indicators of translationese for originally authored and manually translated text classification. We focused on separating originals and translations produced by humans with different levels of expertise and showed that translationese features based on information density and quality estimation are useful indicators of whether a text was manually translated or originally produced. We conducted experiments in a mixed-domain setting, including literary, tourism and scientific texts, as well as instruction manuals, commercial letters and political essays and speeches.

Our experiments on feature type evaluation show that the best performing one is a set of quality estimation inspired distortion indicators, extracted from backward language models trained on originally authored and translated texts. When features are evaluated individually according to the "f-score" measure (Chen and Lin, 2006), the most discriminative ones are from the complexity subset, extracted from n-gram frequency quartiles, followed by surprisal features, both extracted at the lexical and syntactic levels. The features ensemble evaluation based on randomised trees reveals feature complementarity and shows that extracting complexity and distortion indicators at the lexical and POS levels leads to the highest performing sets.

The features used in our experiments are extracted at the word-level. As future work, we plan to extend our feature sets to information theoretic aspects of character-level indicators, such as character n-grams frequencies and language models, encoding complexity and surprisal respectively. This approach would allow to capture sub-word information density indicators, such as morphological information (Avner et al., 2014).

Acknowledgments

This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding[4].

We would like to thank the anonymous reviewers for their insightful comments.

[4]IDEAL – http://www.sfb1102.uni-saarland.de/

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of ACL*, pages 289–295.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2014. Identifying translationese at the word and subword level. *Digital Scholarship in the Humanities*.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In G. Francis Baker M. and E. Tognini-Bonelli, editors, *Text and Technology: in Honour of John Sinclair*, pages 233–250. Benjamins, Amsterdam.

Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. In *JHU/CLSP Summer Workshop Final Report*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of COLING*, pages 315–321.

Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. 2003. Syntactic annotation of a German newspaper corpus. In *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 73–87. Springer.

Michael Carl and Matthias Buch-Kromann. 2010. Correlating translation product and translation process data of professional and student translators. In *Proceedings of EAMT*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Yi-Wei Chen and Chih-Jen Lin. 2006. Combining SVMs with various feature selection strategies. In *Feature extraction*, pages 315–324. Springer.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Monika Doherty. 2006. *Structural propensities: translating nominal word groups from English into German*, volume 65. John Benjamins Publishing.

Jacques Duchateau, Kris Demuynck, and Patrick Wambacq. 2002. Confidence scoring based on backward language models. In *Proceedings of ICASSP*, volume 1.

Birgitta Englund Dimitrova. 2005. *Expertise and explicitation in the translation process*, volume 64. John Benjamins Publishing.

Cathrine Fabricius-Hansen. 1996. Informational density: a problem for translation and translation theory. *Linguistics*, 34(3):521–566.

Austin Frank and T Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the cognitive science society*, pages 933–938.

Simona Gandrabur and George Foster. 2003. Confidence estimation for translation prediction. In *Proceedings of CoNLL*, pages 95–102.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63(1):3–42.

Susanne Göpferich and Riitta Jääskeläinen. 2009. Process research into the development of translation competence: Where are we, and where do we need to go? *Across Languages and Cultures*, 10(2):169–191.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, pages 1–8.

Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425.

Iustina-Narcisa Ilisei. 2012. *A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models*. Ph.D. thesis, University of Wolverhampton.

Riitta Jääskeläinen. 1997. *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating. Doctoral dissertation*. Ph.D. thesis, University of Joensuu, Joensuu.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of ACL*, pages 1318–1326.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT Summit*.

Ekaterina Lapshinova-Koltunski. 2013. VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the Workshop on Building and Using Comparable Corpora*, pages 77–86.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

Gennadi Lembersky. 2013. *The Effect of Translationese on Statistical Machine Translation*. Ph.D. thesis, University of Haifa, Israel.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL: System Demonstrations*, pages 55–60.

Sylwia Ozdowska and Andy Way. 2009. Optimal bilingual data for french-english PB-SMT. In *Proceedings of EAMT*, page 96–103.

Christopher Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC*, pages 825–828.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2015. The Haifa corpus of translationese. *arXiv:1509.03611*.

Anna N Rafferty and Christopher D Manning. 2008. Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46.

Raphael Rubino, Jose G. C. de Souza, Jennifer Foster, and Lucia Specia. 2013a. Topic models for translation quality estimation for gisting purposes. In *Proceedings of MT Summit*, pages 295–302.

Raphael Rubino, Jennifer Foster, Rasoul Samed Zadeh Kaljahi, Johann Roturier, and Fred Hollowood. 2013b. Estimating the quality of translated user-generated content. In *Proceedings of IJCNLP*, pages 14–18.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of EAMT*.

Lucia Specia and Jesús Gimenez. 2010. Combining confidence estimation and reference-based metrics for segment level MT evaluation. In *Proceedings of AMTA*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of ASRU*.

Elke Teich. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Nicola Ueffing and Hermann Ney. 2004. Bayes decision rules and confidence measures for statistical machine translation. *Proceedings of Advances in Natural Language Processing*, pages 70–81.

Nicola Ueffing and Hermann Ney. 2005. Word-level confidence estimation for machine translation using phrase-based translation models. In *Proceedings of EMNLP*, pages 763–770.

Nicola Ueffing, Klaus Macherey, and Hermann Ney. 2003. Confidence measures for statistical machine translation. In *Proceedings of MT Summit*.

Vered Volansky. 2012. The features of translationese. Master's thesis, University of Haifa.