The approximate randomization approach in MultEval (Clark et al., 2011) is used to test whether differences among system performances are statistically significant with a p-value < 0.05.

5 Evaluation of Ontology Labels

In this Section, we report the translation quality of ontology labels based on translation systems learned from different sentence selection methods. Additionally, we perform experiments training an SMT system on the combination of in- and out-domain knowledge. The final approach enhances a domain-specific translation system with lexical knowledge identified in IATE or DBpedia.

5.1 Automatic Translation Evaluation

We report the automatic evaluation based on BLEU and METEOR for the sentence selection techniques, the combination of in- and out-domain data and the lexical enhancement of SMT.

Sentence Selection Techniques As a first evaluation, we automatically compare the quality of the ICD labels translated with different SMT systems trained on specific sentences by the aforementioned selection techniques (Table 2). Due to the in-domain bilingual knowledge, the translation system trained using the EMEA dataset performs slightly better compared to the large generic baseline system. Among the different sentence selection approaches, the *infrequent n-gram recovery* method (infreq. in Table 2) outperforms the baselines and all the other techniques. This is due to the very strict criteria of selecting relevant sentences that allows the *infrequent n-gram recovery* method to identify a limited number (20,000) of highly ontology-specific bilingual sentences. The *related words* and the *n-gram overlap* models perform slightly better than the baseline, with a usage of 81,000 and 59,000 relevant sentences, and perform similarly to the in-domain EMEA translation system.

Dataset Type	Size	BLEU-2	BLEU-4	METEOR
Generic dataset	1.9M	17.2	6.6	24.7
EMEA dataset	1.1M	18.5	7.0	25.8
(1) perplexity	89K	17.5	6.8	24.8
(2) tf-idf	21K	12,6	4.9	18,7
(3) infreq.	20K	19.1	8.1	25.3
(4) related w.	81K	18.9	7.0	25.8
(5) n-gram	59K	17.7	7.1	23.3
(5) ∧ (3)	22K	18.9	8.2*	25.1
(5) ∧ (4)	24K	17.3	7.3	23.9
(3) ∧ (4)	24K	18.4	8.4*	25.5*
(5) ∧ (4) ∧ (3)	30K	**20.1**	**8.9***	**27.2***

Table 2: Automatic translation evaluation on the evaluation dataset of the ICD ontology (Size = amount of selected sentences from the generic parallel corpus. bold results = best performance; *statistically significant compared to baseline)

Further translation quality improvement is possible, if sentence selection methods are combined together (last four rows in Table 2). The cosine similarities of the methods are combined together, whereby new thresholds τ are computed on the *development dataset 1* and applied on the ICD *evaluation dataset*. Each combined method showed improvement compared to the stand-alone method. The best overall performance is obtained when combining the *n-gram overlap*, the semantic *related words* and *infrequent n-gram recovery* methods. With this combination, we reduce the amount of parallel sentences by 98% compared to the generic corpus and significantly outperform the baseline by 2.3 BLEU score points. These two factors confirm the capability of the combined approach of selecting only few ontology-specific bilingual sentences (30,000) that allows the SMT system to identify the correct translations in the target ontology domain. This is due to the fact that the three combined methods are quite complementary. In fact, the *n-gram overlap* method selects a relatively large amount of bilingual sentences with few words in common with the label, the *related words* approach identifies bilingual sentences in the ontology target domain, and the infrequent n-gram recovery technique selects few bilingual sentences with only specific n-grams in common with the labels balancing the effect of the n-gram overlap method.

Combining In- and Out-Domain Data Considering the relatively small amount of parallel data extracted with the sentence selecting methods for the SMT community, we evaluate different approaches that combine a large generic translation model with domain-specific data. For this purpose, we use the sentences selected by the best approach ((5)∧(4)∧(3)) in the previous experiments and combine them with the generic parallel dataset. We evaluate the translation performance when *(i)* concatenating the selected domain-specific parallel dataset with the generic