Ontology Label Translation

Mihael Arcan and Paul Buitelaar

Unit for Natural Language Processing,
Digital Enterprise Research Institute (DERI)
National University of Ireland Galway (NUIG)
Galway, Ireland

{mihael.arcan , paul.buitelaar}@deri.org

Abstract

Our research investigates the translation of ontology labels, which has applications in multilingual knowledge access. Ontologies are often defined only in one language, mostly English. To enable knowledge access across languages, such monolingual ontologies need to be translated into other languages. The primary challenge in ontology label translation is the lack of context, which makes this task rather different than document translation. The core objective therefore, is to provide statistical machine translation (SMT) systems with additional context information. In our approach, we first extend standard SMT by enhancing a translation model with context information that keeps track of surrounding words for each translation. We compute a semantic similarity between the phrase pair context vector from the parallel corpus and a vector of noun phrases that occur in surrounding ontology labels. We applied our approach to the translation of a financial ontology, translating from English to German, using Europarl as parallel corpus. This experiment showed that our approach can provide a slight improvement over standard SMT for this task, without exploiting any additional domain-specific resources.

1 Introduction

The biggest barrier for EU-wide cross-lingual business intelligence is the large number of various languages used by banks or investment firms for their financial reports. In contrast to that, most of the ontologies used for knowledge access are available in English, e.g. the financial ontology FINREP[1] (FINancial REPorting) or COREP[2] (COmmon solvency ratio REPorting). To make the targeted transparency of financial information possible, these ontologies have to be translated first into another language; see also (Declerck et al., 2010). The challenge here lies in translating domain-specific ontology vocabulary, e.g. *Equity-equivalent partner loans*, *Subordinated capital* or *Write-downs of long-term financial assets and securities*.

Since domain-specific parallel corpora for SMT are hardly available, we used a large general parallel corpus, whereby a translation model built by such a resource will tend to translate a segment into the most common word sense. This can be seen for instance when we translate the financial ontology label *Equity-equivalent partner loans* from the German GAAP ontology (cf. Section 3.1). Using a baseline SMT system we get the translation *Gerechtigkeit-gleichwertige Partner Darlehen*. Although this label provides contextual information, *equity* is translated into its general meaning, i.e. *Gerechtigkeit* in the meaning of *justice*, *righteousness* or *fairness*, although *Eigenkapital* would be the preferred translation in the financial domain.

To achieve accurate disambiguation we developed a method using context vectors. We extract semantic information from the ontology, i.e. the vocabulary and relations between labels and compare it with the contextual information extracted from a parallel corpus.

The remainder of the paper is organized as fol-

[1] http://eba.europa.eu/Supervisory-Reporting/FINER.aspx
[2] http://eba.europa.eu/Supervisory-Reporting/COREP.aspx

lows. Section 2 gives an overview of the related work on including semantic information into SMT. Section 3 describes the ontology and the parallel corpus used in our experiment. Then we describe the approach of enhancing the standard SMT model with ontological knowledge for improving the translation of labels in Section 4. In Section 5 the results of exploiting the ontological knowledge described in the previous section are illustrated. Finally we conclude our findings and give an outlook for further research.

2 Related Work

Word sense disambiguation (WSD) systems generally perform on the word level, for an input word they generate the most probable meaning. On the other hand, state of the art translation systems operate on sequences of words. This discrepancy between unigrams versus n-grams was first described in (Carpuat and Wu, 2005). Likewise, (Apidianaki et al., 2012) use a WSD classifier to generate a probability distribution of phrase pairs and to build a local language model. They show that the classifier does not only improve the translation of ambiguous words, but also the translation of neighbour words. We investigate this discrepancy as part of our research in enriching the ontology label translation with ontological information. Similar to their work we incorporate the idea of enriching the translation model with neighbour words information, whereby we extend the window to 5-grams.

(Mauser et al., 2009) generate a lexicon that predicts the bag of output words from the bag of input words. In their research, no alignment between input and output words is used, words are chosen based on the input context. The word predictions of the input sentences are considered as an additional feature that is used in the decoding process. This feature defines a new probability score that favours the translation hypothesis containing words, which were predicted by the lexicon model. Similarly, (Patry and Langlais, 2011) train a model by translating a bag-of-words. In contrast to their work, our approach uses bag-of-word information to enrich the missing contextual information that arises from translating ontology labels in isolation.

(McCrae et al., 2011) exploit in their research the ontology structure for translation of ontologies and taxonomies. They compare the structure of the monolingual ontology to the structure of already translated multilingual ontologies, where the source and target labels are used for the disambiguation process of phrase pairs. We incorporated the idea of using the ontology structure, but avoided the drawback of exploiting existing domain-specific multilingual ontologies.

3 Data sets

For our experiment we used a general parallel corpus to generate the mandatory SMT phrase table and language model. Further, the corpus was used to generate feature vectors on the basis of the contextual information provided by surrounding words. Finally we calculate the semantic similarity between the extracted information from the parallel corpus and the ontology vocabulary.

3.1 Financial ontology

For our experiment we used the financial ontology German GAAP (Generally Accepted Accounting Practice),[3] which holds 2794 concepts with labels in German and English.

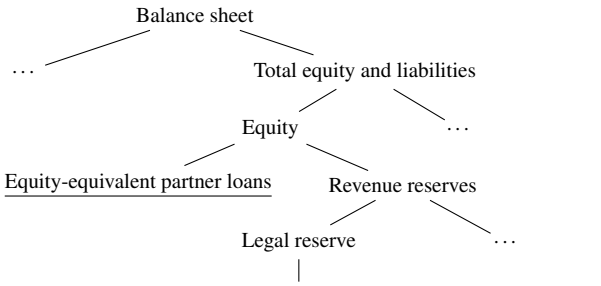

Figure 1: The financial label *Equity-equivalent partner loans* and its neighbours in the German GAAP ontology

As seen in Figure 1 the financial labels do not correspond to phrases from a linguistic point of view. They are used in financial or accounting reports as unique financial expressions or identifiers to organise and retrieve the reported information automatically. Therefore it is important to translate these financial labels with exact meaning preservation.

[3] http://www.xbrl.de/

3.2 Europarl

As a baseline approach we used the Europarl parallel corpus,[4] which holds proceedings of the European Parliament in 21 European languages. We used the English-German parallel corpus with around 1.9 million aligned sentences and 40 million English and 43 million German tokens (Koehn, 2005).

Although previous research showed that a translation model built by using a general parallel corpus cannot be used for domain-specific vocabulary translation (Wu et al., 2008), we decided to train a baseline translation model on this general corpus to illustrate any improvement steps gained by enriching the standard approach with the semantic information of the ontology vocabulary and structure.

4 Experiment

Since ontology labels (or label segments) translated by the Moses toolkit (Section 4.1) do not have much contextual information, we addressed this lack of information and generated fromthe Europarl corpus a new resource with contextual information of surrounding words as feature vectors (Section 4.2). A similar approach was done with the ontology structure and vocabulary (Section 4.3).

4.1 Moses toolkit

To translate the English financial labels into German, we used the statistical translation toolkit Moses (Koehn et al., 2007), where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The SRILM toolkit (Stolcke, 2002) was used to build the 5-gram language model.

4.2 Building the contextual-semantic resource from the parallel corpus Europarl

To enhance the baseline approach with additional semantic information, we built a new resource of contextual information from Europarl.

From the original phrase table, which was generated from the Europarl corpus, we used the sub-phrase table, which was generated to translate the German GAAP financial ontology in the baseline approach. Although this sub-phrase table holds only segments necessary to translate the financial labels, it still contains 2,394,513 phrase pairs. Due to the scalability issue, we reduced the number of phrase pairs by filtering the sub-phrase table based on the following criteria:

a) the direct phrase translation probability $\phi(e|f)$ has to be larger than 0.0001

b) a phrase pair should not start or end with a functional word, i.e. prepositions, conjunctions, modal verbs, pronouns

c) a phrase pair should not start with punctuation

After applying these criteria to the sub-phrase table, the new filtered phrase table holds 53,283 entities, where phrase pairs, e.g. *tax rate* ||| *Steuersatz* or *tax liabilities* ||| *Steuerschulden* were preserved.

In the next step, the phrase pairs stored in the filtered phrase table were used to find sentences in Europarl, where these phrase pairs appear. The goal was to extract the surrounding words as the contextual information of these phrase pairs. If a segment from the filtered phrase table appeared in the sentence we extracted the lemmatised contextual information of the phrase pair, whereby we considered 10 tokens to the left and 10 to the right of the analysed phrase pair. To address the problem of different inflected forms (*financial asset* vs. *financial assets*) of the same lexical entity (*financial asset*) we lemmatised the English part of the Europarl corpus with TreeTagger(Schmid, 1995). Similar to the phrase table filtering approach, an n-gram should not start with a functional word or punctuation. The extracted surrounding words were stored together with its phrase pairs, i.e. for the phrase pairs *Equity-Gerechtigkeit* and *Equity-Eigenkapital* different contextual vectors were generated.

Example 1.a) illustrates a sentence, which holds the source segment *Equity* from the filtered phrase table. Example 1.b) represents its translation into German. This example illustrates the context in which *Equity* is translated into the German expression *Gerechtigkeit*. The segment *Equity* is also present in the second sentence, (example 2.a)), in contrast to the first one, *equity* is translated into *Eigenkapital*, (2.b)), since the sentence reports financial information.

1. a) ... which could guarantee a high standard of efficiency, safety and **equity** for employees and users alike, right away.

[4] http://www.statmt.org/europarl/, version 7

b) ... , der heute ein hohes Niveau an Leistung, Qualität, Sicherheit und **Gerechtigkeit** für die Bediensteten und die Nutzer garantieren könnte.

2. a) ... or organisations from making any finance, such as loans or **equity**, available to named Burmese state-owned enterprises.

   b) ... bzw. Organisationen zu verbieten, birmanischen staatlichen Unternehmen jegliche Finanzmittel wie Darlehen oder **Eigenkapital** zur Verfügung zu stellen.

Applying this methodology on all 1.9 million sentences in Europarl, we generated a resource with feature vectors for all phrase pairs of the filtered phrase table. Table 1 illustrates the contextual differences between the vectors for *Equity-Gerechtigkeit* and *Equity-Eigenkapital* phrase pairs.

4.3 Contextual-semantic resource generation for the financial ontology German GAAP

To compare the contextual information extracted from Europarl a similar approach was applied to the vocabulary in the German GAAP ontology.

First, to avoid unnecessary segments, e.g. *provisions for* or *losses from executory*, we parsed the financial ontology with the Stanford parser (Klein and Manning, 2003) and extracted meaningful segments from the ontology labels. This step was done primarily to avoid comparing all possible n-gram segments with the filtered segments extracted from the Europarl corpus (cf. Subsection 4.2). With the syntactical information given by the Stanford parser we extracted a set of noun segments for the ontology labels, which we defined by the rules shown in Table 2.

#	Syntactic Patterns
1	(NN(S) w+)
2	(NP (NN(S) w+)+))
3	(NP (JJ w+)+ (NN(S) w+)+))
4	(NP (NN(S) w+)+ (CC w+) (NN(S) w+)+)
5	(NP (NN(S) w+)+ (PP (IN/.. w+) (NP (NN(S) w+)+))

Table 2: Syntactic patterns for extracting noun segments from the parsed financial ontology labels

Applying these patterns to the ontology label *Provisions for expected losses from executory contracts* extracts the following noun segments: *provisions*, *losses* and *contracts* (pattern 1), *expected losses* and *executory contracts* (pattern 3), *provisions for expected losses* and *expected losses from executory contracts* (pattern 5).

In the next step, for all 2794 labels from the financial ontology, a unique contextual vector was generated as follows: for the label *Equity-equivalent partner loans* (cf. Figure 1), the vector holds the extracted (lemmatised) noun segments of the direct parent, *Equity*, and all its siblings in the ontology, e.g. *Revenue reserves* ... (Table 3).

targeted label:	Equity-equivalent partner loans
contextual information:	capital (6), reserve (3), loss (3), balance sheet (2) ... currency translation (1), negative consolidation difference (1), profit (1)

Table 3: Contextual information for the financial label *Equity-equivalent partner loans*

4.4 Calculating the Semantic Similarity

Using the resources described in the previous sections in a final step we apply the Cosine, Jaccard and Dice similarity measures on these feature vectors.

For the first evaluation step we translated all financial labels with the general translation model. Table 4 illustrates the translation of the financial expression *equity* as part of the label *Equity-equivalent partner loans*.[5]

With the n-best (n=50) translations for each financial label we calculated the semantic similarity between the contextual information of the phrase pairs (*equity-Eigenkapital*) extracted from the parallel corpus (cf. Table 1) with the semantic information of the financial label *Equity* extracted from the ontology (cf. Table 3).

After calculating a semantic similarity, we reorder the translations based on this additional information, which can be seen in Table 5.

[5] ger. Gerechtigkeit-gleichwertige Partner Darlehen

Source label	Target label	$p(e\|f)$
equity	Gerechtigkeit	-10.6227
equity	Gleichheit	-11.5476
equity	Eigenkapital	-12.7612
equity	Gleichbehandlung	-13.0936
equity	Fairness	-13.6301

Table 4: Top five translations and its translation probabilities generated by the Europarl translation model

Source label	Target label	Context (frequency)
equity	Gerechtigkeit	social (19), efficiency (18), efficiency and equity (14), justice (13), social equity (11), education (9), principle (8), transparency (7), training (7), great (7)
equity	Eigenkapital	capital (19), equity capital (15), venture (3), venture capital (3), rule (2), capital and risk (2), equity capital and risk (2), bank (2), risk (2), debt (1)

Table 1: Contextual information for *Equity* with its target labels *Gerechtigkeit* and *Eigenkapital* extracted from the Europarl corpus

Source label	Target label	Jaccard
equity	Eigenkapital	0.0780169232
equity	Equity	0.0358268041
equity	Kapitalbeteiligung	0.0341965597
equity	Gleichheit	0.0273327211
equity	Gerechtigkeit	0.0266209669

Table 5: Top five re-ranked translations after calculating the Jaccard similarity

5 Evaluation

Our evaluation was conducted on the translations generated by the baseline approach, using only Europarl, and the ontology-enhanced translations of financial labels.

We undertook an automatic evaluation using the BLEU (Papineni et al., 2002), NIST (Doddington, 2002), TER (Snover et al., 2006), and Meteor[6] (Denkowski and Lavie, 2011) algorithms.

5.1 Baseline Evaluation of general corpus

At the beginning of our experiment, we translated the financial labels with the Moses Toolkit, where the translation model was generated from the English-German Europarl aligned corpus. The results are shown in Table 7 as *baseline*.

5.2 Baseline Evaluation of filtered general corpus

A second evaluation on translations was done on a filtered Europarl corpus, depending if a sentence holds the vocabulary of the ontology to be translated. We generated five training sets, based on n-grams of the ontology vocabulary (from unigram to 5-gram) appearing in the sentence. From the set of aligned sentences we generated new translation models and translated again the financial ontology labels with them. Table 6 illustrates the results of filtering the Europarl parallel corpus into smaller (n-gram) training sets, whereby no training set outperforms significantly the baseline approach.

model	sentences	BLEU-4	Meteor	OOV
baseline	1920209	4.22	0.1138	37
unigram	1591520	4.25	0.1144	37
bigram	322607	4.22	0.1077	46
3-gram	76775	1.99	0.0932	92
4-gram	4380	2.45	0.0825	296
5-gram	259	0.69	0.0460	743

Table 6: Evaluation results for the filtered Europarl baseline translation model (OOV - out of vocabulary)

5.3 Evaluation of the knowledge enhanced general translation model

The final part of our research concentrated on translations where the general translation model was enhanced with ontological knowledge. Table 7 illustrates the results using the different similarity measures, i.e. Dice, Jaccard, Cosine similarity coefficient.

For the Cosine coefficient we performed two approaches. For the first step we used only binary values (bv) from the vector, where in the second approach we used the frequencies of the contextual information as real values (rv). The results show that the Cosine measure using frequencies (rv) performs best for the METEOR metric. On the other hand the binary Cosine measure (bv) performs better than the other metrics in BLEU-2 and NIST metrics.

The Jaccard and Dice similarity coefficient perform very similar. They both outperform the general translation model in BLEU, NIST and TER metrics, whereby the Jaccard coefficient performs slightly better than the Dice coefficient. On the other hand both measures perform worse on the METEOR metric regarding the general model. Overall we observe that the Jaccard coefficient outperforms the baseline

[6]Meteor configuration: -l de, exact, stem, paraphrase

	Bleu-2	Bleu-4	NIST	Meteor	TER
baseline	13.05	4.22	1.789	0.113	1.113
Dice	13.16	4.43	1.800	0.111	1.075
Jaccard	13.17	**4.44**	1.802	0.111	**1.074**
Cosine (rv)	12.91	4.20	1.783	**0.117**	1.108
Cosine (bv)	**13.27**	4.34	**1.825**	0.116	1.077

Table 7: Evaluation results for Europarl baseline translation model and the different similarity measures

approach by 0.22 BLEU points.

5.4 Comparison of translations provided by the general model and Jaccard similarity

Table 7 illustrates the different approaches that were performed in our research. As the automatic metrics give just a slight intuition about the improvements of the different approaches, we compared the translations of the general translation model manually with the translations on which Jaccard similarity coefficient was performed.

As discussed, *Equity* can be translated into German as *Gerechtigkeit* when translating it in a general domain or into *Eigenkapital* when translating it in the financial domain. In the financial ontology, the segment *Equity* appears 126 times. The general translation model translates it wrongly as *Gerechtigkeit*, whereby the Jaccard coefficient, with the help of contextual information, favours the preferred translation *Eigenkapital*. Furthermore *Equit* can be also part of a larger financial label, e.g. *Equity-equivalent partner loans*, but the general translation model still translates it into *Gerechtigkeit*. This can be explained by the segmentation during the decoding process, i.e. the SMT system tokenises this label into separate tokens and translates each token separately from each other. On the contrary, the Jaccard similarity coefficient corrects the unigram segment to *Eigenkapital*.

As part of the label *Uncalled unpaid contributions to subscribed capital (deducted from equity on the face of the balance sheet)*, *equity* is again translated by the general translation model as *Gerechtigkeit*. In this case the Jaccard coefficient cannot correct the translation, which is caused by the general model itself, since in all n-best (n=50) translations *equity* is translated as *Gerechtigkeit*. In this case the Jaccard coefficient reordering does not have any affect.

The manual analysis further showed that the ambiguous ontology label *Securities*, e.g. in *Write-downs of long-term financial assets and securities* was also often translated as *Sicherheiten*[7] in the meaning of *certainties* or *safeties*, but was corrected by the Jaccard coefficient into *Wertpapiere*, which is the correct translation in the financial domain.

Finally, the analysis showed that the segment *Balance* in *Central bank balances* was often translated by the baseline model into *Gleichgewichte*,[8] i.e. *Zentralbank Gleichgewichte*, whereas the Jaccard coefficient favoured the preferred translation *Guthaben*, i.e. *Zentralbank Bankguthaben*.

Conclusion and Future Work

Our approach to re-using existing resources showed slight improvements in the translation quality of the financial vocabulary. Although the contextual information favoured correct translations in the financial domain, we see a need for more research on the contextual information stored in the parallel corpus and also in the ontology. Also more work has to be done on analysis of the overlap of the contextual information and the ontology vocabulary, e.g. which contextual words should have more weight for the similarity measure. Furthermore, dealing with the ontology structure, the relations between the labels, i.e. part-of and parent-child relations, have to be considered. Once these questions are answered, the next step will be to compare the classical cosine measure against more sophisticated similarity measures, i.e. Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007). Instead of measuring similarity between the vectors directly using cosine, we will investigate the application of ESA to calculate the similarities between short texts by taking their linguistic variations into account (Aggarwal et al., 2012).

Acknowledgments

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project and Grant No. 296277 for the EuroSentiment project as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

[7] ger. Abschreibungen der langfristigen finanziellen Vermögenswerte und Sicherheiten

[8] en. *equilibrium*, *equation*, *balance*

References

Aggarwal, N., Asooja, K., and Buitelaar, P. (2012). DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. In *SemEval-2012*.

Apidianaki, M., Wisniewski, G., Sokolov, A., Max, A., and Yvon, F. (2012). Wsd for n-best reranking and local language modeling in smt. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9, Jeju, Republic of Korea. Association for Computational Linguistics.

Carpuat, M. and Wu, D. (2005). Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 387–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Declerck, T., Krieger, H.-U., Thomas, S. M., Buitelaar, P., O'Riain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D., and Montiel-Ponsoda, E. (2010). Ontology-based multilingual access to financial reports for sharing business knowledge across europe. In *Internal Financial Control Assessment Applying Multilingual Ontology Framework*.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceeding of the 41st annual meeting of the association for computational linguistics*, pages 423–430.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL*, ACL '07, pages 177–180.

Mauser, A., Hasan, S., and Ney, H. (2009). Extending statistical machine translation with discriminative and trigger-based lexicon models. In *Conference on Empirical Methods in Natural Language Processing*, pages 210–217, Singapore.

McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., and Cimiano, P. (2011). Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5)*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patry, A. and Langlais, P. (2011). Going beyond word cooccurrences in global lexical selection for statistical machine translation using a multilayer perceptron. In *5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 658–666, Chiang Mai, Thailand.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

Wu, H., Wang, H., and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 993–1000.