Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 214–220,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

Learning automatic metrics in a pairwise setting, i.e., learning to distinguish between two alternative translations and to decide which of the two is better (which is arguably one of the easiest ways to produce a ranking), emulates closely how human judges perform evaluation assessments in reality. Instead of learning a similarity function between a translation and the reference, they learn how to differentiate a better from a worse translation given a corresponding reference. While the pairwise setting does not provide an absolute quality scoring metric, it is useful for most evaluation and MT development scenarios.

In this paper, we propose a pairwise learning setting similar to that of Duh (2008), but we extend it to a new level, both in terms of feature representation and learning framework. First, we integrate several layers of linguistic information encapsulated in tree-based structures; Duh (2008) only used lexical and POS matches as features. Second, we use information about both the reference and two alternative translations *simultaneously*, thus bringing our ranking closer to how humans rank translations. Finally, instead of deciding upfront which types of features between hypotheses and references are important, we use a our structural kernel learning (SKL) framework to generate and select them automatically.

The structural kernel learning (SKL) framework we propose consists in: (*i*) designing a structural representation, e.g., using syntactic and discourse trees of translation hypotheses and a references; and (*ii*) applying structural kernels (Moschitti, 2006; Moschitti, 2008), to such representations in order to automatically inject structural features in the preference re-ranking algorithm. We use this method with translation-reference pairs to directly learn the features themselves, instead of learning the importance of a predetermined set of features. A similar learning framework has been proven to be effective for question answering (Moschitti et al., 2007), and textual entailment recognition (Zanzotto and Moschitti, 2006).

Our goals are twofold: (*i*) in the short term, to demonstrate that structural kernel learning is suitable for this task, and can effectively learn to rank hypotheses at the segment-level; and (*ii*) in the long term, to show that this approach provides a unified framework that allows to integrate several layers of linguistic analysis and information and to improve over the state-of-the-art.

Below we report the results of some initial experiments using syntactic and discourse structures. We show that learning in the proposed framework yields better correlation with humans than applying the traditional translation–reference similarity metrics using the same type of structures. We also show that the contributions of syntax and discourse information are cumulative. Finally, despite the limited information we use, we achieve correlation at the segment level that outperforms BLEU and other metrics at WMT12, e.g., our metric would have been ranked higher in terms of correlation with human judgments compared to TER, NIST, and BLEU in the WMT12 Metrics shared task (Callison-Burch et al., 2012).

2 Kernel-based Learning from Linguistic Structures

In our pairwise setting, each sentence s in the source language is represented by a tuple $\langle t_1, t_2, r\rangle$, where t_1 and t_2 are two alternative translations and r is a reference translation. Our goal is to develop a classifier of such tuples that decides whether t_1 is a better translation than t_2 given the reference r.

Engineering features for deciding whether t_1 is a better translation than t_2 is a difficult task. Thus, we rely on the automatic feature extraction enabled by the SKL framework, and our task is reduced to choosing: (*i*) a meaningful structural representation for $\langle t_1, t_2, r\rangle$, and (*ii*) a feature function ϕ_{mt} that maps such structures to substructures, i.e., our feature space. Since the design of ϕ_{mt} is complex, we use tree kernels applied to two simpler structural mappings $\phi_M(t_1, r)$ and $\phi_M(t_2, r)$. The latter generate the tree representations for the translation-reference pairs (t_1, r) and (t_2, r). The next section shows such mappings.

2.1 Representations

To represent a translation-reference pair (t, r), we adopt shallow syntactic trees combined with RST-style discourse trees. Shallow trees have been successfully used for question answering (Severyn and Moschitti, 2012) and semantic textual similarity (Severyn et al., 2013b); while discourse information has proved useful in MT evaluation (Guzmán et al., 2014; Joty et al., 2014). Combined shallow syntax and discourse trees worked well for concept segmentation and labeling (Saleh et al., 2014a).

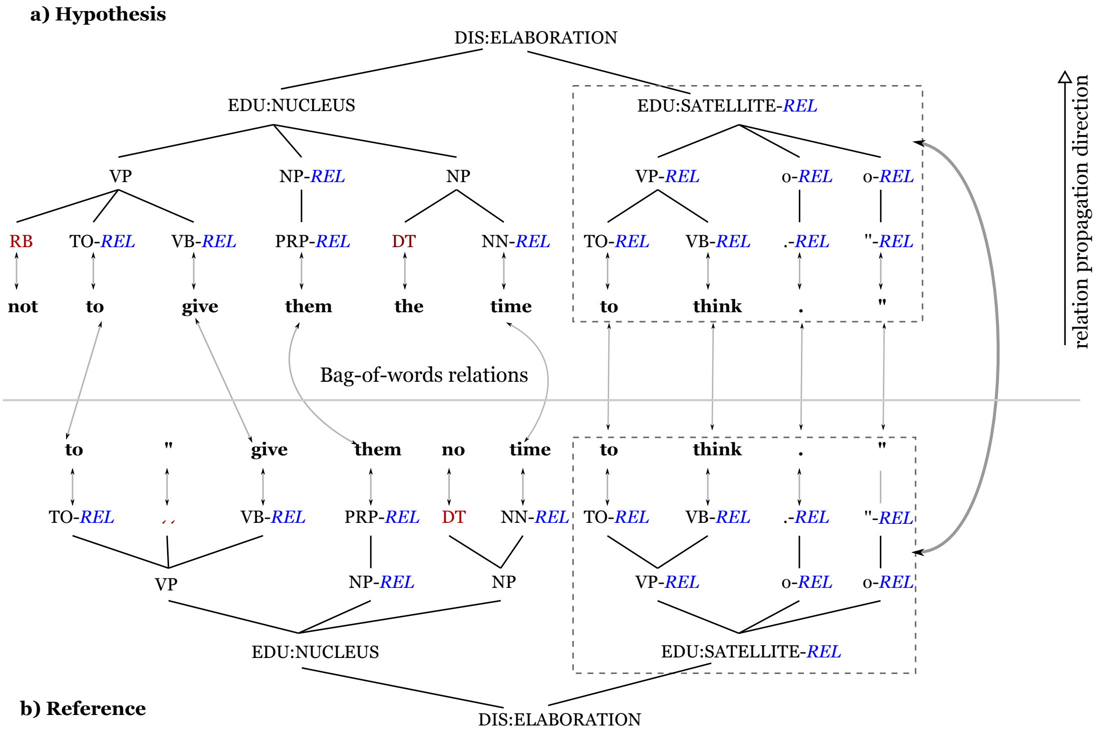

Figure 1: Hypothesis and reference trees combining discourse, shallow syntax and POS.

Figure 1 shows two example trees combining discourse, shallow syntax and POS: one for a translation hypothesis (top) and the other one for the reference (bottom). To build such structures, we used the Stanford POS tagger (Toutanova et al., 2003), the Illinois chunker (Punyakanok and Roth, 2001), and the discourse parser[1] of (Joty et al., 2012; Joty et al., 2013).

The lexical items constitute the leaves of the tree. The words are connected to their respective POS tags, which are in turn grouped into chunks. Then, the chunks are grouped into elementary discourse units (EDU), to which the nuclearity status is attached (i.e., NUCLEUS or SATELLITE). Finally, EDUs and higher-order discourse units are connected by discourse relations (e.g., DIS:ELABORATION).

2.2 Kernels-based modeling

In the SKL framework, the *learning objects* are pairs of translations $\langle t_1, t_2\rangle$. Our objective is to automatically learn which pair features are important, independently of the source sentence. We achieve this by using kernel machines (KMs) over two learning objects $\langle t_1, t_2\rangle, \langle t_1', t_2'\rangle$, along with an explicit and structural representation of the pairs (see Fig. 1).

More specifically, KMs carry out learning using the scalar product

$$K_{mt}(\langle t_1, t_2\rangle, \langle t_1', t_2'\rangle) = \phi_{mt}(t_1, t_2) \cdot \phi_{mt}(t_1', t_2'),$$

where ϕ_{mt} maps pairs into the feature space.

Considering that our task is to decide whether t_1 is better than t_2, we can conveniently represent the vector for the pair in terms of the difference between the two translation vectors, i.e., $\phi_{mt}(t_1, t_2) = \phi_K(t_1) - \phi_K(t_2)$. We can approximate K_{mt} with a preference kernel PK to compute this difference in the kernel space K:

$$\begin{aligned} & PK(\langle t_1, t_2\rangle, \langle t_1', t_2'\rangle) \qquad (1) \\ & = K(t_1) - \phi_K(t_2)) \cdot (\phi_K(t_1') - \phi_K(t_2')) \\ & = K(t_1, t_1') + K(t_2, t_2') - K(t_1, t_2') - K(t_2, t_1') \end{aligned}$$

The advantage of this is that now $K(t_i, t_j') = \phi_K(t_i) \cdot \phi_K(t_j')$ is defined between two translations only, and not between two pairs of translations. This simplification enables us to map translations into simple trees, e.g., those in Figure 1, and then to apply them tree kernels, e.g., the Partial Tree Kernel (Moschitti, 2006), which carry out a scalar product in the subtree space.

We can further enrich the representation ϕ_K, if we consider all the information available to the human judges when they are ranking translations. That is, the two alternative translations along with their corresponding reference.

[1] The discourse parser can be downloaded from http://alt.qcri.org/tools/

In particular, let r and r' be the references for the pairs $\langle t_1, t_2\rangle$ and $\langle t_1', t_2'\rangle$, we can redefine all the members of Eq. 1, e.g., $K(t_1, t_1')$ becomes

$$K(\langle t_1, r\rangle, \langle t_1', r'\rangle) = \text{PTK}(\phi_M(t_1, r), \phi_M(t_1', r')) + \text{PTK}(\phi_M(r, t_1), \phi_M(r', t_1')),$$

where ϕ_M maps a pair of texts to a single tree.

There are several options to produce the bitext-to-tree mapping for ϕ_M. A simple approach is to only use the tree corresponding to the first argument of ϕ_M. This leads to the basic model $K(\langle t_1, r\rangle, \langle t_1', r'\rangle) = \text{PTK}(\phi_M(t_1), \phi_M(t_1')) + \text{PTK}(\phi_M(r), \phi_M(r'))$, i.e., the sum of two tree kernels applied to the trees constructed by ϕ_M (we previously informally mentioned it).

However, this simple mapping may be ineffective since the trees within a pair, e.g., (t_1, r), are treated independently, and no meaningful features connecting t_1 and r can be derived from their tree fragments. Therefore, we model $\phi_M(r, t_1)$ by using word-matching *relations* between t_1 and r, such that connections between words and constituents of the two trees are established using position-independent word matching. For example, in Figure 1, the thin dashed arrows show the links connecting the matching words between t_1 and r. The propagation of these relations works from the bottom up. Thus, if all children in a constituent have a link, their parent is also linked.

The use of such connections is essential as it enables the comparison of the structural properties and relations between two translation-reference pairs. For example, the tree fragment [ELABORATION [SATELLITE]] from the translation is connected to [ELABORATION [SATELLITE]] in the reference, indicating a link between two entire discourse units (drawn with a thicker arrow), and providing some reliability to the translation[2].

Note that the use of connections yields a graph representation instead of a tree. This is problematic as effective models for graph kernels, which would be a natural fit to this problem, are not currently available for exploiting linguistic information. Thus, we simply use K, as defined above, where the mapping $\phi_M(t_1, r)$ only produces a tree for t_1 annotated with the marker REL representing the connections to r. This marker is placed on all node labels of the tree generated from t_1 that match labels from the tree generated from r.

In other words, we only consider the trees enriched by markers separately, and ignore the edges connecting both trees.

3 Experiments and Discussion

We experimented with datasets of segment-level human rankings of system outputs from the WMT11 and the WMT12 Metrics shared tasks (Callison-Burch et al., 2011; Callison-Burch et al., 2012): we used the WMT11 dataset for training and the WMT12 dataset for testing. We focused on translating into English only, for which the datasets can be split by source language: Czech (cs), German (de), Spanish (es), and French (fr). There were about 10,000 non-tied human judgments per language pair per dataset. We scored our pairwise system predictions with respect to the WMT12 human judgments using the Kendall's Tau (τ), which was official at WMT12.

Table 1 presents the τ scores for all metric variants introduced in this paper: for the individual language pairs and overall. The left-hand side of the table shows the results when using as similarity the direct kernel calculation between the corresponding structures of the candidate translation and the reference[3], e.g., as in (Guzmán et al., 2014; Joty et al., 2014). The right-hand side contains the results for structured kernel learning.

We can make the following observations:

(*i*) The overall results for all SKL-trained metrics are higher than the ones when applying direct similarity, showing that learning tree structures is better than just calculating similarity.

(*ii*) Regarding the linguistic representation, we see that, when learning tree structures, syntactic and discourse-based trees yield similar improvements with a slight advantage for the former. More interestingly, when both structures are put together in a combined tree, the improvement is cumulative and yields the best results by a sizable margin. This provides positive evidence towards our goal of a unified tree-based representation with multiple layers of linguistic information.

(*iii*) Comparing to the best evaluation metrics that participated in the WMT12 Metrics shared task, we find that our approach is competitive and would have been ranked among the top 3 participants.

[2] Note that a non-pairwise model, i.e., $K(t_1, r)$, could also be used to match the structural information above, but it would not learn to compare it to a second pair (t_2, r).

[3] Applying tree kernels between the members of a pair to generate one feature (for each different kernel function) has become a standard practice in text similarity tasks (Severyn et al., 2013b) and in question answering (Severyn et al., 2013a).

		Similarity					Structured Kernel Learning				
	Structure	*cs*-en	*de*-en	*es*-en	*fr*-en	all	*cs*-en	*de*-en	*es*-en	*fr*-en	all
1	Syn	0.169	0.188	0.203	0.222	0.195	0.190	0.244	0.198	0.158	0.198
2	Dis	0.130	0.174	0.188	0.169	0.165	0.176	0.235	0.166	0.160	0.184
3	Dis+Pos	0.135	0.186	0.190	0.178	0.172	0.167	0.232	0.202	0.133	0.183
4	Dis+Syn	0.156	0.205	0.206	0.203	0.192	**0.210**	**0.251**	**0.240**	**0.223**	**0.231**

Table 1: Kendall's (τ) correlation with human judgements on WMT12 for each language pair.

Furthermore, our result (0.237) is ahead of the correlation obtained by popular metrics such as TER (0.217), NIST (0.214) and BLEU (0.185) at WMT12. This is very encouraging and shows the potential of our new proposal.

In this paper, we have presented only the first exploratory results. Our approach can be easily extended with richer linguistic structures and further combined with some of the already existing strong evaluation metrics.

		Testing				
	Train	*cs*-en	*de*-en	*es*-en	*fr*-en	all
1	*cs*-en	0.210	0.204	0.217	0.204	0.209
2	*de*-en	0.196	0.251	0.203	0.202	0.213
3	*es*-en	0.218	0.204	**0.240**	0.223	0.221
4	*fr*-en	0.203	0.218	0.224	0.223	0.217
5	all	**0.231**	**0.258**	0.226	**0.232**	**0.237**

Table 2: Kendall's (τ) on WMT12 for cross-language training with Dis+Syn.

Note that the results in Table 1 were for training on WMT11 and testing on WMT12 for each language pair in isolation. Next, we study the impact of the choice of training language pair. Table 2 shows cross-language evaluation results for Dis+Syn: lines 1-4 show results when training on WMT11 for one language pair, and then testing for each language pair of WMT12.

We can see that the overall differences in performance (see the last column: *all*) when training on different source languages are rather small, ranging from 0.209 to 0.221, which suggests that our approach is quite independent of the source language used for training. Still, looking at individual test languages, we can see that for de-en and es-en, it is best to train on the same language; this also holds for fr-en, but there it is equally good to train on es-en. Interestingly, training on es-en improves a bit for cs-en.

These somewhat mixed results have motivated us to try tuning on the full WMT11 dataset; as line 5 shows, this yielded improvements for all language pairs except for es-en. Comparing to line 4 in Table 1, we see that the overall Tau improved from 0.231 to 0.237.

4 Conclusions and Future Work

We have presented a pairwise learning-to-rank approach to MT evaluation, which learns to differentiate good from bad translations in the context of a given reference. We have integrated several layers of linguistic information (lexical, syntactic and discourse) in tree-based structures, and we have used the structured kernel learning to identify relevant features and learn pairwise rankers.

The evaluation results have shown that learning in the proposed SKL framework is possible, yielding better correlation (Kendall's τ) with human judgments than computing the direct kernel similarity between translation and reference, over the same type of structures. We have also shown that the contributions of syntax and discourse information are cumulative, indicating that this learning framework can be appropriate for the combination of different sources of information. Finally, despite the limited information we used, we achieved better correlation at the segment level than BLEU and other metrics in the WMT12 Metrics task.

In the future, we plan to work towards our long-term goal, i.e., including more linguistic information in the SKL framework and showing that this can help. This would also include more semantic information, e.g., in the form of Brown clusters or using semantic similarity between the words composing the structure calculated with latent semantic analysis (Saleh et al., 2014b).

We further want to show that the proposed framework is flexible and can include information in the form of quality scores predicted by other evaluation metrics, for which a vector of features would be combined with the structured kernel.

Acknowledgments

This research is part of the Interactive sYstems for Answer Search (Iyas) project, conducted by the Arabic Language Technologies (ALT) group at Qatar Computing Research Institute (QCRI) within the Qatar Foundation.

References

Joshua Albrecht and Rebecca Hwa. 2008. Regression for machine translation evaluation at the sentence level. *Machine Translation*, 22(1-2):1–27.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 136–158, Prague, Czech Republic.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 22–64, Edinburgh, Scotland, UK.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 10–51, Montréal, Canada.

Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 333–338, Uppsala, Sweden.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, WMT '08, pages 191–194, Columbus, Ohio, USA.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 256–264, Prague, Czech Republic.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 687–698, Baltimore, Maryland, USA.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 904–915, Jeju Island, Korea.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 486–496, Sofia, Bulgaria.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, pages 402–408, Baltimore, Maryland, USA.

Alon Lavie and Michael Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, Michigan, USA.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 243–252, Montréal, Canada.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 776–783, Prague, Czech Republic.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, ECML/PKDD '06, pages 318–329, Berlin, Germany.

Alessandro Moschitti. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 253–262, Napa Valley, California, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meting of the Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, USA.

Maja Popović and Hermann Ney. 2007. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, WMT '07, pages 48–55, Prague, Czech Republic.

Vasin Punyakanok and Dan Roth. 2001. The use of classifiers in sequential inference. In *Advances in Neural Information Processing Systems 14*, NIPS '01, pages 995–1001, Vancouver, Canada.

Iman Saleh, Scott Cyphers, Jim Glass, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2014a. A study of using syntactic and semantic structures for concept segmentation and labeling. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 193–202, Dublin, Ireland.

Iman Saleh, Alessandro Moschitti, Preslav Nakov, Lluís Màrquez, and Shafiq Joty. 2014b. Semantic kernels for semantic parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, Doha, Qatar.

Aliaksei Severyn and Alessandro Moschitti. 2012. Structural relationships for large-scale learning of answer re-ranking. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 741–750, Portland, Oregon, USA.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013a. Learning adaptable patterns for passage reranking. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, CoNLL '13, pages 75–83, Sofia, Bulgaria.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013b. Learning semantic textual similarity with structural representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '13, pages 714–718, Sofia, Bulgaria.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, AMTA '06, Cambridge, Massachusetts, USA.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence-level MT evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 123–129, Edinburgh, Scotland, UK.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, HLT-NAACL '03, pages 173–180, Edmonton, Canada.

Billy Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1060–1068, Jeju Island, Korea.

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, COLING-ACL '06, pages 401–408, Sydney, Australia.

Differentiation refers to a wide variety of teaching techniques and lesson adaptations that educators use to instruct a diverse group of students, with diverse learning needs, in the same course, classroom, or learning environment. Differentiation is commonly used in â€œheterogeneous groupingâ€â€”an educational strategy in which students of different abilities, learning needs, and levels of academic achievement are grouped together. In heterogeneously grouped classrooms, for example, teachers vary instructional strategies and use more flexibly designed lessons to engage student interests and addres The needs of the learner were being better catered for, but the teacher was up all night. She needed to think about differentiation in a different way. 10 ways to differentiate learningâ€¦ 1. Let go. Give the students (at least some) ownership of their learning. Donâ€™t always be the boss of the class, be part of the community of learners. Donâ€™t make all the decisions. Allow choice. Encourage students to think about how they learn best. Have students decide how to demonstrate their learning. 2. Change your expectations. We present a pairwise learning-to-rank approach to machine translation evalua-tion that learns to differentiate better from worse translations in the context of a given reference. We integrate several layers of linguistic information encapsulated in tree-based structures, making use of both the reference and the system output simul-taneously, thus bringing our ranking closer to how humans evaluate translations. Most importantly, instead of deciding upfront which types of features are important, we use the learning framework of preference re-ranking kernels to learn the features au-tomatically.