Semantic Textual Similarity for MT evaluation

Julio Castillo[†‡] **Paula Estrella**[‡]
[‡]FaMAF, UNC, Argentina
[†‡]UTN-FRC, Argentina
jotacastillo@gmail.com
pestrella@famaf.unc.edu.ar

Abstract

This paper describes the system used for our participation in the WMT12 Machine Translation evaluation shared task.
We also present a new approach to Machine Translation evaluation based on the recently defined task Semantic Textual Similarity. This problem is addressed using a textual entailment engine entirely based on WordNet semantic features.
We described results for the Spanish-English, Czech-English and German-English language pairs according to our submission on the Eight Workshop on Statistical Machine Translation.
Our first experiments reports a competitive score to system level.

1 Introduction

The evaluation of Machine Translation (MT) has become as important as MT itself over the last few years. This is evidenced by the fact that there are now specific forums to present and test new metrics, such as the Workshop for Statistical MT (WMT) or the NIST MetricsMatr. Every year a vast number of MT metrics are created, the majority being automatic, and seeking to find an efficient, low labor-intensive and reliable evaluation method as an alternative to human-based evaluation.

Automatic metrics employ different evaluation strategies: classical MT automatic metrics, such as BLEU (Papineni et al., 2002), NIST (Doddington. 2002), WER (Tillmann et al., 1997), PER (Nießen et al., 2000) are language-independent based on n-gram matching (considering or not the ordering of words in a sentence); other use some kind of language-specific knowledge, for example METEOR (Banerjee et al., 2005), which uses WordNet to match synonyms if exact matchings do not occur, and METEOR-NEXT (Denkowski et al., 2010) that, in addition to METEOR's features, incorporates paraphrases; and more sophisticated metrics use deeper linguistic information, as for example the DCU-LFG metric (Yifan et al., 2010).

However, relatively few attempts have been made to use semantic information for MT evaluation. Moreover, only one work has been published about using semantic equivalence (known as Textual Entailment) of texts for MT evaluation. In this work we propose an improved metric, based on TE features, that indicates to what extent a candidate sentence is equivalent to a reference.

The paper is organized as follows: Section 2 describes the relevant work done on semantic oriented MT evaluation, Section 3 describes the architecture of the system to compute our metric, then Section 4 relates TE and semantic textual similarity to MT, and Section 5 presents some results obtained with our TE-based metric; and finally Section 6 summarize some conclusions and future work.

2 Related work

Given the vast literature in the field of MT evaluation, in this section we briefly mention a few attempts to evaluate MT based on semantic features, which we deem most recent and important.

2.1 Semantics for MT evaluation

Giménez and Márquez (2007) present a set of metrics operating over shallow semantic structures, which they call linguistic elements, with the idea that a sentence can be seen as a 'bag' of LEs. Possible LEs are word forms, part-of-speech tags, dependency relationships, syntactic phrases, named

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 52–58,
Montréal, Canada, June 7-8, 2012.

entities, semantic roles, etc. The metrics calculate the similarity of a candidate to one or more references by calculating the overlap and matches of LEs, and the resulting score is the highest obtained from the individual comparisons to each reference. The shallow-semantic evaluation is performed by computing the matching and overlap of named entities and semantic roles, after automatically annotating the sentences.

Following this work, Giménez and Márquez (2009) propose the family of metrics discourse representation structure (DRS) based on the Discourse Representation Theory of Kamp (1981), where a discourse is represented in structure that is essentially a variation of first-order predicate calculus. These sets of metrics are then used to evaluate poor quality MT, concluding that semantic oriented metrics are more stable at the system level, while at the sentence level their performance decreases (probably due to external factors, for example if a parse tree of the sentence is not available, the metric cannot be computed).

More recently, Lo and Wu (2011) present a new semi-automated metric, MEANT, that assesses translation utility by matching semantic role fillers. Their hypothesis is that a good translation is one that lets a reader get the central information of the sentence. Conceptually, MEANT is defined in terms of f-score, calculated by averaging the translation accuracy for all frames in the MT output across the number of frames in the MT output/reference translations. To determine the translation accuracy for each semantic role filler in the reference and machine translations, they ask humans to indicate if a role filler translation is correct, incorrect or partially correct, hence being a semi-automatic metric. According to Lo and Wu (2011) MEANT can be run using inexpensive untrained monolingual human judges and yet it correlates with human judgments on adequacy as well as other labor-intensive metrics, such as HTER (Snover et al., 2006), which needs to train humans to find the closest right translation.

2.2 Textual Entailment in MT

Textual Entailment (TE) is defined as a generic framework for applied semantic inference, where the core task is to determine whether the meaning of a target textual assertion (hypothesis, H) can be inferred from a given text (T). For example, given the pair (H,T):

H: The Tunisian embassy in Switzerland was attacked
T: Fire bombs were thrown at the Tunisian embassy in Bern
we can conclude that T entails H.

The recently created challenge "Recognising Textual Entailment" (RTE) started in 2005 with goal of providing a binary answer for each pair (H,T), namely whether there is entailment holds or not (Dagan et al., 2006). The RTE challenge has mutated over the years, aiming at accomplishing more accurate and specific solutions; for example, 2008 a three-way decision was proposed (instead of the original binary decision) consisting of "entailment", "contradiction" and "unknown"; in 2009 the organizers proposed a pilot task, the Textual Entailment Search (Bentivogli et al., 2009), consisting in finding all the sentences in a set of documents that entail a given Hypothesis and since 2010 there is a Novelty Detection Task, which means that RTE systems are required to judge whether the information contained in each H is novel with respect to (i.e., not entailed by) the information contained in the corpus.

This task is quite close to the goal of MT and MT evaluation given that a correct translation should be semantically equivalent to its reference, and thus both translations should entail each other.

Despite this close relation, at present there are only two works using TE in MT, namely Mirkin et al. (2009) proposes to handle OOV(Out-of-vocabulary words) terms by generating alternative source sentences for translation but instead of simply using paraphrases they use entailed texts; the other contribution is by Aziz et al. (2010), in which TE features are integrated into standard SMT workflow (i.e. they dynamically generate alternative entailed words to replace OOVs).

More directly related to our work, is that of Padó et al., (2009) that uses TE to evaluate MT. The main idea is to find out if the translation paraphrases (entails) the reference using entailment features. This is implementing by checking for entailment both from the candidate to the reference and from the reference to the candidate; best candidates are thus assumed to be those that both entail and are entailed by the references and worst candidates are assumed to be those that neither entail the references nor are entailed by these references. Padó et al. (2009a) found that entailment-

based features extracted from partially ill-formed translations are sufficiently robust to be predictive for translation quality.

Our approach differs from that of Padó et al. (2009) in that we do not have a binary entailment relation; instead we try to state in a scale of 0 – 5 the degree of similarity between a candidate and a reference. This approach has very recently been proposed as a new task of the Semantic Evaluation Exercises 2012, called Semantic Textual Similarity (STS) by Aguirre et al. (2012) and is explained in more detail in Section 4.

3 System architecture

Sagan is a RTE textual entailment system which has taken part of several challenges, including the Textual Analysis Conference 2009 and TAC 2010, and the Semantic Textual Similarity (Castillo and Estrella, 2012) and Cross Lingual Textual Entailment for content synchronization (Castillo and Cardenas, 2012) as part of the *SEM 2012 Task8 (Negri et al., 2012).

The system is based on a machine learning approach for STS. We adapted this system to produce feature vectors for all MT outputs for all language pairs ES-EN, DE-EN, FR-EN and CS-EN. It is worth noting that we work on all pairs into English because the system was run in a monolingual setting to take advantage of all the resources available for EN.

This Semantic Textual Similarity engine utilizes eight WordNet-based similarity measures, as explained in (Castillo, 2011), with the purpose of obtaining the maximum similarity between two concepts. These text-to-text similarity measures are based on the following word-to-word similarity metrics: (Resnik, 1995), (Lin, 1997), (Jiang and Conrath, 1997), (Pirrò and Seco, 2008), (Wu & Palmer, 1994), Path Metric, (Leacock & Chodorow, 1998), and a semantic similarity to sentence level named SemSim (Castillo and Cardenas, 2010).

Additional information about how to produce feature vector and metric to word and sentence level can be found in (Castillo, 2011).

The output of the system as modified for this workshop, is a similarity score between 5 and 0, where 5 means a perfect semantic similarity (applied to MT it means that a candidate is indeed a good translation) and 0 means that there is no semantic similarity between the pair, i.e. in MT terms, the candidate is not a translation.

The architecture of the system is shown in Figure 1.

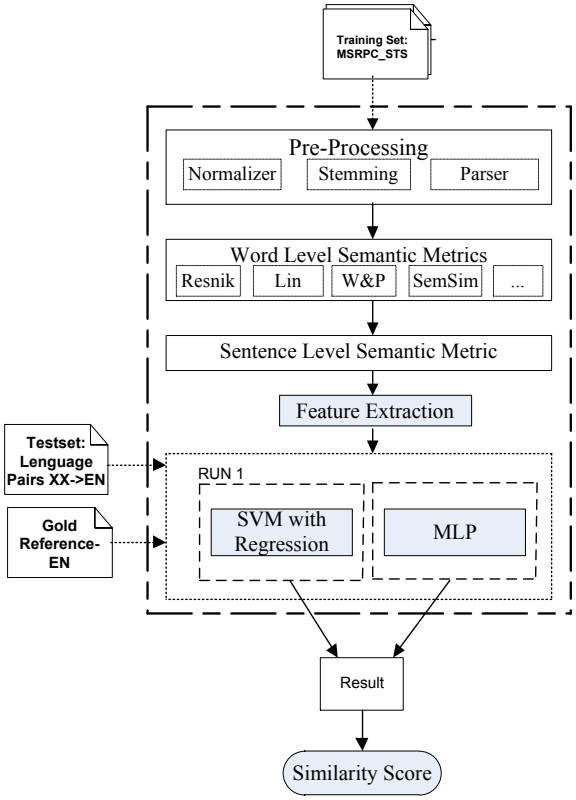

Fig.1. STS system architecture for MT evaluation

The system computes the semantic similarity of two texts (T,H) as a function of the semantic similarity of the constituent words of both phrases. A graph matching algorithm is used to determine the overall similarity between two text fragments.

As a result, a text to text similarity measure is built based on word to word similarity. It is assumed that combining word to word similarity metrics to text level would be a good indicator of text to text similarity.

4 Sagan for MT evaluation

Sagan for MT evaluation is based on a core development to approach the Semantic Textual Similarity task(STS). The pilot task STS was recently defined in Semeval 2012 (Aguirre et al., 2012) and has as main objective measuring the degree of semantic equivalence between two text fragments. STS is related to both Recognizing Textual Entailment (RTE) and Paraphrase Recognition, but

has the advantage of being a more suitable model for multiple NLP applications.

As mentioned before, the goal of the RTE task (Bentivogli et al., 2009) is determining whether the meaning of a hypothesis H can be inferred from a text T. Thus, TE is a directional task and we say that T entails H, if a person reading T would infer that H is most likely true. The difference with STS is that STS consists in determining how similar two text fragments are, in a range from 5 (total semantic equivalence) to 0 (no relation). Thus, STS mainly differs from TE and Paraphrasing in that the classification is graded instead of binary. In this manner, STS is filling the gap between TE and Paraphrase.

In view of this, our claim is that the output of MT systems will be more strongly correlated with humans if we have a higher STS score between MT system output and the reference translation.

To apply Sagan to MT evaluation, we first, pre-process the pairs from Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005) with dates and time normalization, and then optional modules are applied depending on the metric we want to calculate. Second, we compute 8 sentence level semantic features, and, finally, for every segment generated by systems participating at WMT 2012, we determine the semantic similarity score between that output and the given reference translation. The scores are then normalized to a value in the range 0 – 1.

5 Experiments and results

For the WMT 2012 we participated in the Czech-English and Spanish-English evaluation task but we did not have enough time to extensively test our metric on a diverse range of settings (i.e. different corpora and language pairs), given that it was developed for the STS task, which released the data and results only a couple of months ago.

However, we are now running experiments to get a better picture of the metric's ability to rate translation quality. In this section we report results obtained by training the system on the WMT 2011 data and testing on the news test portion, only for the Spanish-English pair. Although the system handles both SVM with regression and MLP classifiers, well known to have good performance on natural language applications, we only submit the results obtained using SVM with regression due to previous experiments that consistently showed higher accuracy using SVM instead of MLP.

At the system level, we calculated the Spearman Rank Correlation Coefficient (ρ) to compare our metric's behavior with respect to the human based metric applied in WMT 2011. The result is $\rho = 0.96$ indicating a strong positive correlation. Moreover, we successfully reproduce the systems ranking given by humans regarding the best and worst systems.

System Id	Human score	Sagan score
online-B	0.72	0.71
online-A	0.72	0.71
systran	0.66	0.7
koc	0.67	0.69
alacant	0.66	0.69
rbmt-1	0.63	0.69
rbmt-4	0.6	0.69
rbmt-3	0.61	0.69
uedin	0.51	0.68
rbmt-2	0.6	0.68
upm	0.5	0.68
rbmt-5	0.51	0.68
ufal-um	0.47	0.67
cu-zeman	0.16	0.59
hyderabad	0.17	0.58

Table 1. Sagan's score for ES-EN WMT 2011 news test set.

When correlating our metric to other automatic metrics, we find that it better correlates with Meteor-Rank and Adq (Denkowski and Lavie, 2011), Tesla-b (Dahlmaier et al., 2011) and MPF (Popovic, 2011), with a correlation coefficient of 0.96. On the other hand, the worst correlations are found against Tesla-f, F15 (Bicici and Yuret, 2011) and the TER baseline (Snover et al., 2006).

We also performed experiments to segment level with the language pair ES-EN. We used the MSR_STS as training set and the newstest2011 from WMT 2011 as test set. MSR_STS[1] is composed by 750 sentence pairs with a graded semantic relationship ranging from 5 (equivalence) to 0 (no-equivalence).

As result, we obtained a Kendall-tau correlation coefficient of 0.29 to segment-level for translations

[1] http://www.cs.york.ac.uk/semeval-2012/task6/

into English. These preliminary results, although low, shows that STS and Textual Entailment could be used to address the problem of MT evaluation. Clearly, further improvements are needed and we suspect that higher score can be reached using bigger training data. We also remark the necessity of larger corpus of STS providing a graded score among sentences.

At the segment level, we show in Table 2 some examples found by manually inspecting the results.

Example Number	MT output	Texts	Sagan score
2397	Reference	Adelaida, 4 years old, wants a doll or a bicycle, while her sister Isabel, 3 years old, would like a Barbie doll.	0.95
	Online-A	Adelaide, of 4 years, want a doll or a bicycle, while **his** sister Isabel, 3 years, would like a Barbie doll.	
2417	Reference	"I strongly rely on the Charter."	0.18
	Online-A	"Me I based mainly on the letter."	
45	Reference	But there is a snag in that.	0.105
	Alacant	However, there is a fly in the ointment.	
1510	Reference	Unfortunately, even Scarlett Johansson might struggle to raise China's subterranean regard for these city squads.	0.5206
	cu-zeman	**Lamentablemente**, until scarlett johansson should fight to increase the **ínfimo** respect of china for with these **escuadrones** the city.	

Table 2. Sagan's score for some illustrative ES-EN WMT 2011 example pairs showing the score between MT outputs and the reference translation.

The example number 2397 shows a sentence that achieves a high score (0.95) but that has an agreement error (marked in bold), that prevented Sagan from assigning the highest score.

Otherwise, the instance number 2417 has a score of 0.18 showing that Sagan correctly penalizes ill-formed or meaningless sentences. Similarly, the example number 45 has a very low score which quantifies the dissimilarity with the reference translation.

Finally, the last example provided shows that the translation remains words in the original Spanish language (marked in bold).

This manual inspection will be complemented with a deeper study of the correlations at the sentence level.

6 Conclusions and future work

In this paper we introduced a new metric for MT evaluation based on Semantic Textual Similarity computed over textual entailment features. The metric's goal is to provide an indicative score of the extent to which two texts (a candidate translation and a reference) are equivalent. This goal is more complex than classical binary decisions in the field of TE and is a new approach to bring together the knowledge from different areas that a similar ambitions.

While promising results were found at the system level, the metric still needs to be tested on a diversity of settings and at the segment level; this is work in progress and results will be reported in due time.

References

Jesús Giménez and Lluís Márquez. 2007. *Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems*. In Proceedings of the ACL Workshop on Statistical Machine Translation, pages 256–264.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. *Accelerated DP Based Search For Statistical Translation*. In Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation*. In Proceedings of the 40th AnnualMeeting of the Association for Computational Linguistics(ACL-02), pages 311–318.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. *A Evaluation Tool for Machine Translation:Fast Evaluation for MT Research*. In

Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000).

G. Doddington. 2002. *Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics.* In Proceedings of the 2nd International Conference on HLT, pp. 138–145, San Francisco, CA, USA.

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In Proceedings of the 43th ACL, pages 65–72.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR.

He Yifan, Du Jinhua, Way Andy, and Van Josef . 2010. *The DCU dependency-based metric in WMT-MetricsMATR 2010.* In: WMT 2010 - Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, ACL, Uppsala, Sweden.

Kamp H. 1981. *A theory of truth and semantic representation*. In Groenendijk, J., Janssen, T., & Stokhof, M. (Eds.), Formal methods in the study of language, No. 135, pp. 277–322. Mathematical Centre, Amsterdam.

Chi-kiu Lo and Dekai Wu. 2011. *MEANT: inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles*. 49th Annual Meeting of the Association for Computational Linguistic (ACL-2011). Portland, Oregon, US.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. *A Study of Translation Edit Rate with Targeted Human Annotation*. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06), pages 223–231.

Ido Dagan, Oren Glickman and Bernardo Magnini. 2006. *The PASCAL Recognising Textual Entailment Challenge*. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.) Machine Learning Challenges. LNCS, Vol. 3944, pp. 177-190.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. *Source-language entailment modeling for translating unknown terms*. ACL 2009. Vol. 2. Stroudsburg, PA, USA, 791-799.

Wilker Aziz and Marc Dymetmany and Shachar Mirkin and Lucia Specia and Nicola Cancedda and Ido Dagan. 2010. *Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-VocabularyWords*. In: Proceedings of the 14th annual meeting of the European Association for Machine Translation (EAMT), Saint-Rapha, France.

Dahlmeier, Daniel and Liu, Chang and Ng, Hwee Tou. 2011.TESLA at WMT 2011: Translation Evaluation and Tunable Metric.In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, pages 78-84, Edinburgh, Scotland.

S. Pado, D. Cer, M. Galley, D. Jurafsky and C. Manning. 2009. *Measuring Machine Translation Quality as Semantic Equivalence: A Metric Based on Entailment Features*. Journal of MT 23(2-3), 181-193.

S. Pado, M. Galley, D. Jurafsky and C. Manning. 2009a. *Robust Machine Translation Evaluation with Entailment Features*. Proceedings of ACL 2009.

Eneko Agirre, Daniel Cer, Mona Diab and Aitor Gonzalez-Agirre. 2012. *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity*. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Bentivogli, Luisa, Dagan Ido, Dang Hoa, Giampiccolo, Danilo, Magnini Bernardo.2009.*The Fifth PASCAL RTE Challenge*. In: Proceedings of the TAC.

Castillo Julio. 2011. *A WordNet-based semantic approach to textual entailment and cross-lingual textual entailment*. International Journal of Machine Learning and Cybernetics - Springer, Volume 2, Number 3.

Estrella Paula, Popescu-Belis A. and King M. 2007. *A New Method for the Study of Correlations between MT Evaluation Metrics and Some Surprising Results.* In: Proceedings of TMI-07- 11th Conference on Theoretical and Methodological Issues in Machine Translation -, Skvvde, Sweden.

Castillo Julio and Cardenas Marina. 2010. *Using sentence semantic similarity based onWordNet in recognizing textual entailment*. Iberamia 2010. In LNCS, vol 6433. Springer, Heidelberg, pp 366–375.

Castillo Julio. 2010. *A semantic oriented approach to textual entailment using WordNet-based measures*. MICAI 2010. LNCS, vol 6437. Springer, Heidelberg, pp 44–55.

Castillo Julio. 2010. *Using machine translation systems to expand a corpus in textual entailment*. In: Proceedings of the Icetal 2010. LNCS, vol 6233, pp 97–102.

Resnik P. 1995. *Information content to evaluate semantic similarity in a taxonomy*. In: Proceedings of IJCAI 1995, pp 448–453 907

Lin D. 1997.*An information-theoretic definition of similarity*. In: Proceedings of Conference on Machine Learning, pp 296–304 909

Jiang J, Conrath D.1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In: Proceedings of theROCLINGX 911

Pirro G., Seco N. 2008. *Design, implementation and evaluation of a new similarity metric combining fea-*

ture and intrinsic information content. In: ODBASE 2008, Springer LNCS.

Wu Z, Palmer M. 1994. *Verb semantics and lexical selection*. In: Proceedings of the 32nd ACL 916.

Leacock C, Chodorow M. 1998. *Combining local context and WordNet similarity for word sense identification*. MIT Press, pp 265–283 919

Hirst G, St-Onge D . 1998. *Lexical chains as representations of context for the detection and correction of malapropisms*. MIT Press, pp 305–332 922

Banerjee S, Pedersen T. 2002. *An adapted lesk algorithm for word sense disambiguation using WordNet*. In: Proceeding of CICLING-02

Castillo Julio and Estrella Paula. 2012. *SAGAN: An approach to Semantic Textual Similarity based on Textual Entailment.* In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Castillo Julio and Cardenas Marina. 2012. *SAGAN: A Machine Translation Approach for Cross-Lingual Textual Entailment*. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Chris Callison-Burch, Philipp Koehn, Christof Monz, Omar Zaidan. 2011.*Findings of the 2011Workshop on Statistical Machine Translation*. WMT 2011.

M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. 2012. Semeval-2012. *Task 8: Cross-lingual Textual Entailment for Content Synchronization*. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012).

William B. Dolan and Chris Brockett.2005. *Automatically Constructing a Corpus of Sentential Paraphrases*. Third International Workshop on Paraphrasing (IWP2005). Asia Federation of Natural Language Processing.