COGEX at RTE3

Marta Tatu and **Dan Moldovan**
Language Computer Corporation
Richardson, Texas, 75080
United States
marta,moldovan@languagecomputer.com

Abstract

This paper reports on LCC's participation at the Third PASCAL Recognizing Textual Entailment Challenge. First, we summarize our semantic logical-based approach which proved successful in the previous two challenges. Then we highlight this year's innovations which contributed to an overall accuracy of 72.25% for the RTE 3 test data. The novelties include new resources, such as eXtended WordNet KB which provides a large number of world knowledge axioms, event and temporal information provided by the TARSQI toolkit, logic form representations of events, negation, coreference and context, and new improvements of lexical chain axiom generation. Finally, the system's performance and error analysis are discussed.

1 Introduction

Continuing a two-year tradition, the PASCAL Network organized the Third Recognizing Textual Entailment Challenge[1] (RTE 3) to further the research on reasoning systems able to decide whether the meaning of one text (the entailed hypothesis, H) can be inferred from another text (the entailing text, T). Among this year's challenges, approximately 15% of the (T, H) pairs contained long texts (more details in Section 5.1).

We approach the *textual entailment* problem as a *logical implication between meanings* (Fowler et al., 2005; Tatu et al., 2006). Our system transforms the two text snippets into three-layered semantically-rich logic form representations, generates an abundant set of lexical, syntactic, semantic, and world knowledge axioms and, iteratively, searches for a proof for the entailment between the text T and a possibly relaxed version of the hypothesis H. A pair is labeled as positive if the score of the found proof (reflecting H's degree of relaxation) is above a threshold learned on the training data. Figure 1 summarizes our approach to RTE.

2 Cogex's Innovations for RTE 3

2.1 EXtended WordNet Knowledge Base

eXtended WordNet Knowledge Base (XWN-KB) is the result of our ongoing research which captures and stores the rich world knowledge encoded in WordNet's glosses into a knowledge base. In XWN-KB, the glosses have been transformed into a set of semantic relations using a semantic parser whose output has been verified by human annotators. Fig. 2 displays the semantic relations derived for *Nobel laureate*'s definition. Our system used this representation for QA Dev pair 579 and QA Test pair 582 [2].

2.2 TARSQI Toolkit

The TARSQI project (Temporal Awareness and Reasoning Systems for Question Interpretation)[3] (Verhagen et al., 2005) builds a modular system which detects, resolves and normalizes time expressions (both absolute and relative times) - GUTime tagger; marks events and their grammatical features -

[1] www.pascal-network.org/Challenges/RTE3

[2] Table 6 lists the pairs referenced throughout the paper.

[3] http://www.timeml.org/site/tarsqi

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 22–27,
Prague, June 2007. ©2007 Association for Computational Linguistics

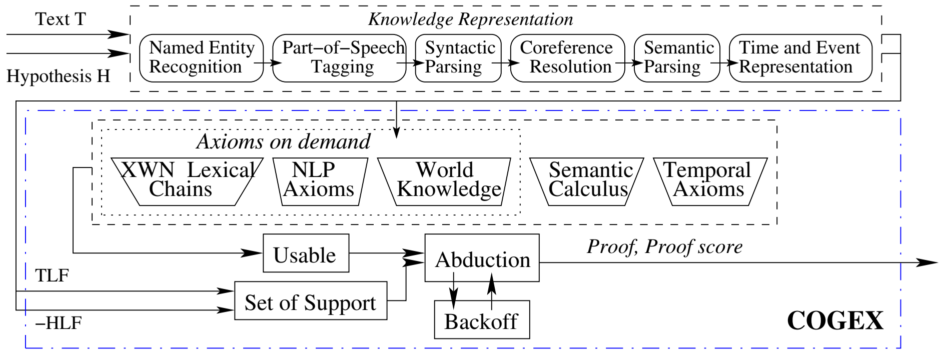

Figure 1: Cogex's Architecture

The Pet passport alone can be [used]$_{e_1:occurrence}$ to [enter]$_{e_2:occurrence}$ the UK, but it will not [suffice]$_{e_3:occurrence}$ to [enter]$_{e_4:occurrence}$ many countries. For instance Guatemala, like almost every country, [demands]$_{e_5:occurrence}$ that all imported pets have a rabies vaccination, but will not [accept]$_{e_6:i\_action}$ the Pet passport as proof of [said]$_{e_7:reporting}$ vaccination. **modality**: (e_1:*can*); **tense**: (e_2:infinitive), (e_3:future), (e_4:infinitive), (e_5:present), (e_6:future), (e_7:past); **polarity**: (e_3:negative), (e_6:negative); **slink**: modal(e_1, e_2), modal(e_5, e_6), factive(e_6, e_7); **tlink**: before(e_1, e_2), before(e_5, e_6), before(e_4, e_5), before(e_6, e_7), before(e_2, e_3), before(e_2, e_4)

Table 1: TARSQI's Treatment of IE Dev pair 63's T

Nobelist, Nobel_laureate; synsetId: 06822770
gloss: *winner of a Nobel_prize*
ISA(Nobel_laureate,winner) THEME(Nobel_prize,winner)

Figure 2: XWN-KB Treatment of *Nobel_laureate*

Evita; identifies subordination constructions introducing modality information - Slinket; adds temporal relations between events and temporal expressions - GUTenLINK; and computes temporal closures - SputLink. We used the information provided by the TARSQI toolkit (*Run #1*) as an alternative to our event detection and temporal expression identification and normalization modules (*Run #2*). Table 1 shows TARSQI's output for IE Dev pair 63's T.

The following sections present innovations related to the logic form knowledge representation.

2.3 Logic Representation of Events

For events, the logic representation of their describing concept was augmented with a special predicate (`event_EV(e1)`). When we made use of TARSQI's output (Run #1), the event predicate was replaced by the class of the event (`occurrence_EV(e1)`, `state_EV(e1)`, `reporting_EV(e1)`, etc.).

2.4 Negation

Recently, the logic representation of sentences with negated concepts was altered to mark as negated the entire scope of the negation. For example, the logic form of IE Dev pair 90's H: *Kennon did not participate in the WWII*, formerly equal to `Kennon_NN(x1) & -participate_VB(e1,x1,x4) & in_IN(e1,x2) & WWII_NN(x2) & _conflict_NE(x2) & AGT_SR(x1,e1)`, became `Kennon_NN(x1) & WWII_NN(x2) & _conflict_NE(x2) & -(exists e1 (participate_VB(e1,x1,x3) & in_IN(e1,x2) & AGT_SR(x1,e1)))` which is closer to the meaning of the English text snippet. For Run #1 (with TARSQI output), we only used the *polarity* information attached to the identified events and negated the event's predicate.

2.5 Coreference Resolution

In order to cope with the long text pairs, we added in our processing pipeline a dedicated pronominal coreference resolution module which replaced the inter-sentential resolution processing we used until now. The new tool combines Hobbs algorithm (Hobbs, 1978) and the Resolution of Anaphora Procedure (RAP) algorithm (Lappin and Leass, 1994). For the RTE task, it is very important to have tight connections between the predicates of

Semantic Relation	Axiom Templates
ISA	`n1(x1) -> n2(x1);v1(e1,x1,x2) -> v2(e1,x1,x2)`
DERIVATION	`n(x1) -> v(e1,x1,x2) & AGENT_SR(x1,e1);n(e1) -> v(e1,x1,x2)` `v(e1,x1,x2) -> n(x1);v(e1,x1,x2) -> n(e1)`
CAUSE	`v1(e1,x1,x2) -> v2(e2,x2,x3) & CAUSE_SR(e1,e2)`
AGENT	`n1(x1) -> n2(x2) & AGENT_SR(x1,x2)`
PERTAIN	`a(x1,x2) -> n(x1)`

Table 2: Semantic Relation - Axiom Template mapping

long texts. For example, for QA Dev pair 409, resolving the pronoun *he* to *George H.W. Bush* is a step needed to correctly label the pair. But IE Dev pair 92 requires more advanced anaphora resolution which corefers *the team* and *the Kinston Indians*.

3 Natural Language Axiom Improvements

3.1 XWN Lexical Chains

In order to take advantage of XWN-KB, we implemented few changes in our lexical chain axioms generation module. The most significant refinement is the one axiom-per-chain relation approach. Previously, the system was generating one axiom for the entire lexical chain, but, given the diversity of semantic relations which link the WordNet concepts and the difficulty to reduce an entire semantically rich chain to one implication which captures its meaning, a remodeling of our axiom generation module was required. Therefore, for each relation in the best lexical chain found between one of T's constituents and one of H's constituents, an axiom is created. For each semantic relation, we created a set of axiom templates to be used during the axiom generation process. Several examples of axiom templates are shown in Table 2. Therefore, a lexical chain is broken down into several axioms whose relations are combined by the logic prover as it sees fit. For instance, the chain *oil_company#n#1* $\xrightarrow{agent}$ *sell#v#1* $\xleftarrow{entailment}$ *trade#v#1* is translated into the axioms `oil_company_NN(x1) -> sell_VB(e1,x1,x2) & AGENT_SR(x1,e1)` and `sell_VB(e1,x1,x2) -> trade_VB(e1,x1,x2)` used to prove the entailment for IE Dev pair 196.

We also changed the subset of senses considered when lexical chains are built. Previously, this subset contained the first k ($k = 3$) senses for each content word. For this year's challenge, we changed the sense selection mechanism and we used the cluster of WordNet senses to which the fine-grained sense assigned by the Word Sense Disambiguation system corresponds. We used the coarse-grained sense inventory for WordNet 2.1 released for Task #7 in SemEval-2007[4]. This clustering was created automatically with the aid of a methodology described in (Navigli, 2006). For example, the 10 WordNet senses for the noun *bank* are mapped into 3 clusters.

3.2 NLP Axioms

In addition to the syntactic re-writing rules which break down complex syntactic structures, including complex nominals and coordinating conjunctions, we added a new type of NLP axioms which links a named entity to its set of aliases. For IE Dev pair 35, the link between the *Central Intelligence Agency* mentioned in T and H's *CIA* is very important.

We also added a deeper analysis of multi-word human named entities which marks *last names* (*Hawking*), *first (male/female) names* (*Stephen*), *titles* (*Prime Minister*) and *names* for human entities found in WordNet (*Tony Blair*). This fine classification has three goals: (1) to mark human entities with the gender information (used by the pronominal coreference module); (2) to prevent lexical chains to use first names of human entities as their source or target (*Elizabeth* as part of *Elizabeth Alexandra Mary* should not be mapped to {*Elizabeth#1, Elizabeth_II#1*} or {*Elizabeth#2, Elizabeth_I#1*} - QA Dev pair 407); (3) to create more precise NLP axioms for human entities denoting noun compounds. These axioms follow rules such as `title(x1) & last_name(x2) & nn_NNC(x3,x1,x2) -> last_name(x3) & title(x3)`, `title(x1) & first_name(x2) & last_name(x3) & nn_NNC(x4,x1,x2,x3) -> nn_NNC(x4,x2,x3)`, etc. For IR Dev pair 287, the

[4] `nlp.cs.swarthmore.edu/semeval`

axiom `Prime_Minister_NN(x6) & Giulio_NN(x7) & Andreotti_NN(x8) & nn_NNC(x9,x6,x7,x8) & _human_NE(x9) -> Andreotti_NN(x9)` expresses the equivalence between *Prime Minister Giulio Andreotti* and *Andreotti*. During the processing of the development set, the prover used 75 axioms of this type. During testing, 112 axioms proved to be useful in finding proofs.

4 Named Entity Check

Based on the guidelines for judging whether T entails or not H, hypotheses that introduce entities which cannot be derived from T are not entailed by the text (the pair is labeled as *NO*). Therefore, we created a proof's score adjustment module which deducts points for each pair whose H contains at least one *named entity* not-derivable from T. Once the prover used the loaded axioms to derive all the possible information from the text, this named entity check is performed. We note that the named entity heuristic is not equivalent with the removal of a named entity predicate from the hypothesis in the relaxation stage which can also occur if the syntactic constraints in which the named entity participates are not satisfied. For instance, for IR Dev pair 387, *Puncheon Lama* is a new entity introduced by the hypothesis without any connection to the text.

5 Experiments and Results

5.1 Experimental Data

The RTE 3 data set was derived with four NLP applications in mind: Information Extraction (IE), Information Retrieval (IR), Question Answering (QA), and Multi-document Summarization (SUM). Statistics for this year's dataset are shown in Table 3. On average, the *long* texts contain twice the number of words found in texts from pairs marked as *short*.

Dataset	**True**	**False**	**Overall**
IE	105 (8)	95 (11)	200 (19)
IR	87 (23)	113 (31)	200 (54)
QA	106 (22)	94 (13)	200 (35)
SUM	112 (5)	88 (4)	200 (9)
Test	410 (58)	390 (59)	800 (117)
Development	412 (78)	388 (57)	800 (135)

Table 3: Data split between *true* and *false* classes. The number of pairs with *long* text is shown in parenthesis.

5.2 Cogex's Performance

Table 4 details our submission results for Run #1 (TARSQI's events, temporal expressions and event-event and event-time relations) and Run #2 (LCC's event, temporal expressions and event-time relations)[5]. The two runs do not differ significantly. The extra information captured in the logic representations used in Run #1 (as compared with Run #2) was not the focus of the entailment; the understanding it brings was not exercised during the entailment recognition process. For the IR and QA tasks, Run #1 results are better when compared to Run #2's. For these tasks, the performance of the system is much higher when compared with the results obtained for IE and SUM. Even tough the thresholds learned for these two tasks best separate the positive from the negative pairs on the development set, they prove to be fairly low for the test set. Almost all positive IE and SUM pairs are identified as such (very high recall for both tasks), but a lot of negatives are also labeled as positives (low precision, smaller accuracy).

Task	**A**	**AvgP**	**P**	**R**	**F**
Run #1					
IE	63.50	61.44	59.20	98.10	73.84
IR	78.00	78.83	76.54	71.26	73.81
QA	87.50	87.81	87.85	88.68	88.26
SUM	60.00	61.54	58.99	93.75	72.41
Test	**72.25**	69.42	67.41	88.78	76.63
Dev	76.37	72.12	75.17	80.82	77.89
Run #2					
IE	64.50	56.26	60.12	96.19	73.99
IR	75.50	77.65	75.00	65.52	69.94
QA	85.00	87.40	81.67	92.45	86.73
SUM	62.00	58.16	60.47	92.86	73.12
Test	**71.75**	67.97	67.16	87.80	76.11
Dev	74.12	71.28	71.95	81.55	76.45

Table 4: Results for Run #1 and Run #2

5.3 Named Entity Heuristic Impact

Table 5 details the interaction between the prover (Run #1) and the named entity heuristic[6]. The

[5] **A**, **AvgP**, **P**, **R** and **F** stand for accuracy, average precision, precision, recall, and f-measure, respectively.

[6] **Coverage** shows the number of pairs for which the heuristic fired, **H's A**, **C's A** and **C+H's A** indicate, respectively, the named entity heuristic accuracy, Cogex's and the prover's when

heuristic fires for 167 pairs, while only 154 of them are true negative entailments (92.21% accuracy). The prover's accuracy for the same subset of pairs is 65.86%. The maximum overall improvement in accuracy that the heuristic can bring is 5.5%, but, because of the way the heuristic penalizes the proof scores, its overall improvement is 4.12%.

Task	Coverage	H's A	C's A	C+H's A
IE	19	100.00	42.10	100.00
IR	56	94.64	73.21	94.64
QA	59	94.91	77.96	94.91
SUM	33	78.78	45.45	45.45
Test	167	92.21	65.86	85.63

Table 5: NE heuristic's performance for Run #1

In theory, the named entity check should not fail. But, in practice, its performance is influenced by the knowledge that the prover collects and, if this information is not complete, then the heuristic fails. For example, for QA Dev pair 419, H mentions number *three* and because the prover cannot infer it as the cardinality of the elementary particles mentioned in T, the heuristic fires incorrectly.

5.4 Error Analysis

Some of the sources of errors are:

Lexical chains For IR Test pair 377, *black_plague* can be derived from T's *plague* only if we allow lexical chains with more than 2 HYPONYMY relations (*plague#n#1* $\xrightarrow{hyponymy}$ *bubonic_plague#n#1* $\xrightarrow{hyponymy}$ *black_plague#n#1*). This restriction on lexical chains was added last year. However, in this year's data this restriction was detrimental as shown in the above example.

Named entity heuristic Some of the errors introduced by the named entity heuristic are debatable. For example, IR Test pair 355's hypothesis introduces the named entity *German* which cannot be derived from the text. Similarly, for QA Test pairs 495 and 496, the name *Christian Democratic Union* cannot be inferred from the text's mention of *Christian Democrat party*. On the other hand, pairs for which the score adjustment introduced by the named entity heuristic did not change the label assigned by the prover include SUM Test pair 656 whose hypothesis mentions *US* without it being derivable from the text (unless we consider the adjective *domestic*).

World Knowledge For SUM Test pair 744, the system fails to infer *nearly half a million dollars* from $480,350. Similarly, the system failed to entail *died in 1970* from the biographical markings "(1890-1970)" for QA Test pair 486.

High word overlap SUM pairs have a high degree of word-overlap between T and H and detection of the non-entailment requires careful processing. SUM Test pair 666's text contains an extra adverbial phrase which changes the label of the pair.

Reports and Modality Even though reporting verbs (*X said that Y*) and modalities (*X may Y*, *X tried to Y*) should influence the validity of the statement they modify, most Y clauses are considered true in the RTE data (SUM Dev pair 756, IR Dev pair 295, IE Dev pair 148 are just few examples). Therefore, our solutions for representing[7] or checking these modifiers[8] failed to bring any improvement on the development set and were not included in the processing of the test set.

But, for IE Test pair 172, T's main verb is qualified by *threatened* which is not present in H. For SUM Test pair 672, *cited strong volume gains* does not entail *makes strong profits*.

6 Conclusion

The XWN-KB is an invaluable resource for recognizing textual entailment. Its impact in RTE 3 was significant. However, we are still exploring ways of fully exploiting this resource. The use of the TARSQI toolkit did not impact the performance because the temporal knowledge was not exercised in this year's task. Contrary to our expectations, the representation of modality had a negative impact on the performance. This is perhaps due to incorrect representation. For our system, the introduction of long texts did not cause significant problems. The system is robust enough to handle longer texts.

the scores it computes are adjusted according to the heuristic.

[7]Our representation for *X said Y* which prevents the entailment that *Y* is `(X(x1) & _report_CTXT(c1,x1)) -> (Y(e1))`

[8]We attempted to penalize proofs which infer the second argument of an MODAL slink without entailing the first. Pairs IE Dev 191 and IR Dev 203 fall in this category, but have different gold annotation labels.

Id	Tag	Pair Text and Hypothesis
D35	YES	T: A leading human rights group on Wednesday identified Poland and Romania as the likely locations in eastern Europe of secret prisons where al-Qaeda suspects are interrogated by the Central Intelligence Agency. H: CIA secret prisons were located in Eastern Europe.
D92	YES	T: The Kinston Indians are a minor league baseball team in Kinston, North Carolina. The team, a Class A affiliate of the Cleveland Indians, plays in the Carolina League. H: Kinston Indians participate in the Carolina League.
D191	YES	T: Though Wilkins and his family settled quickly in Italy, it wasn't a successful era for Milan, and Wilkins was allowed to leave in 1987 to join French outfit Paris Saint-Germain. H: Wilkins departed Milan in 1987.
D196	YES	T: Some large Russian oil companies, including Lukoil, Zarubezhneft, the state-owned oil company, and Alpha Eco, the trader, were implicated by the report. H: Zarubezhneft trades in oil.
D203	NO	T: A decision to allow the exiled Italian royal family to return to Italy may be granted amid the discovery that the head of the family, Prince Vittorio Emmanuele, addressed the president of Italy properly. He has called President Ciampi "our president, the president of all Italians". H: Italian royal family returns home.
D287	YES	T: Italy's highest court has upheld a court verdict that partially cleared former Prime Minister Giulio Andreotti of having colluded with the Mafia. H: Andreotti is accused of Mafia collusion.
D387	NO	T: Thus, China's President repeatedly sent letters and envoys to the Dalai Lama and to the Tibetan Government asking that Tibet "join" the Republic of China. H: Dalai Lama and the government of the People's Republic of China are in dispute over Panchen Lama's reincarnation.
D409	YES	T: George H.W. Bush served this country not only as President but also as Vice President, Member of Congress, United Nations Ambassador, chief of the U.S. Liaison Office to the People's Republic of China, Director of the Central Intelligence Agency and also, as a naval aviator in World War II. Coming back from the war, he married his sweetheart, Barbara Pierce of Rye, New York, and later that year made his first civilian adult decision when he made the appropriate choice of moving to Texas, where he lived the rest of his life. H: The name of George H.W. Bush's wife is Barbara.
D419	YES	T: Discovery of the top quark, if confirmed, completes one set of subatomic building blocks whose existence is predicted by the prevailing theory, called the Standard Model, of the particles and forces that determine the fundamental nature of matter and energy. In the whimsical lexicon of modern physics, the elementary particles are called quarks, leptons and bosons. H: Quarks, leptons, and bosons are the three elementary particles of physics according to the Standard Model.
D579	YES	T: Salma Hayek drew a crowd in Veracruz, Mexico, at the July 8 premiere of 'Nobody Writes to the Colonel', a movie based on a short novel by Nobel laureate Gabriel Garcia Marquez. H: Gabriel Garcia Marquez is a Nobel prize winner.
D756	YES	T: The contaminated pills included metal fragments ranging in size from "microdots" to portions of wire one-third of an inch long, the FDA said. H: The contaminated pills contained metal fragments.
T172	NO	T: This year thousands of Hindu Holy Men, also known as sadhus, threatened to boycott festivals during their pilgrimage to the Ganges, where their rituals involve washing away their sins by bathing in the water. H: Hindu Holy Men boycotted festivals during their pilgrimage to the Ganges.
T355	YES	T: Before reconstruction began, the Reichstag was wrapped by the Bulgarian artist Christo and his wife Jeanne-Claude in 1995, attracting millions of visitors. H: Christo wraps German Reichstag.
T377	YES	T: The U.S. enjoyed miraculously long immunity from the dreaded plague that used to sweep Europe. H: Black plague swept Europe.
T495	YES	T: Former German Chancellor Helmut Kohl said Thursday he will not break pledges he made to campaign contributors by publicly disclosing their names even though his Christian Democrat party has directed him to reveal their identities. H: The name of Helmut Kohl's political party is the Christian Democratic Union.
T561	YES	T: Pope John Paul II arrived Saturday in the birthplace of the Solidarity movement that he sparked with his first papal visit 20 years ago, offering Poles "the greeting of a fellow Pole who comes among you to fulfill the needs of his own heart." H: Pope John Paul II was born in Poland.
T565	NO	T: On the domestic tobacco front, operating income rose by 12 per cent to Dollars 914m, with "slightly higher unit volume". H: US tobacco income has risen.
T666	NO	T: Boys and girls will be segregated during sex education in junior high school. H: Boys and girls will be segregated in junior high school.
T672	NO	T: Philip Morris cited strong volume gains in Germany, Italy, France, Spain, central and eastern Europe, the Far East, Japan, Korea, Argentina and Brazil. H: Philip Morris makes strong profits also in Europe.
T725	YES	T: After years of battling between oil companies, the Ecuadorian government decided to collaborate with indigenous groups. H: The Ecuadorian government collaborated with indigenous groups.

Table 6: Examples of RTE 3 pairs. D# and T# refer to the # pair from the dev and the test set, respectively

Acknowledgments

We would like to thank Iulia Nica for her helpful suggestions (including the named entity check) and Marc Verhagen and James Pustejovsky for presenting us with TARSQI's output for the RTE datasets.

References

A. Fowler, B. Hauser, D. Hodges, I. Niles, A. Novischi, and J. Stephan. 2005. Applying COGEX to Recognize Textual Entailment. In *Proceedings of the PASCAL Challenges Workshop*.

J. Hobbs. 1978. Resolving Pronoun References. *Lingua*, 44:311–338.

S. Lappin and H. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

R. Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of COLING-ACL 2006*.

M. Tatu, B. Iles, J. Slavick, A. Novischi, and D. Moldovan. 2006. COGEX at the Second Recognizing Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

M. Verhagen, I. Mani, R. Sauri, R. Knippen, J .Littman, and J. Pustejovsky. 2005. Automating Temporal Annotation with TARSQI. In *Proceedings of ACL 2005. Demo Session*.