

**INTERNATIONAL CONFERENCE**

**RECENT ADVANCES IN**

**NATURAL LANGUAGE PROCESSING**

**PROCEEDINGS**

Edited by  
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov

Borovets, Bulgaria  
14-16 September 2009

# TABLE OF CONTENTS

Naveed AFZAL and Viktor PEKAR <i>Unsupervised Relation Extraction for Automatic Generation of Multiple-Choice Questions</i> .....	1
Ahmet AKER and Robert GAIZAUSKAS <i>Summary Generation for Toponym-Referenced Images using Object Type Language Models</i> .....	6
Vincent Van ASCH and Walter DAELEMANS <i>Prepositional Phrase Attachment in Shallow Parsing</i> .....	12
Alexandra BALAHUR, Ester BOLDRINI, Andrés MONTOYO and Patricio MARTÍNEZ-BARCO <i>A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries</i> .....	18
Eduard BARBU <i>Acquisition of Common Sense Knowledge for Basic Level Concepts</i> .....	23
Eduard BARBU and Massimo POESIO <i>Unsupervised Knowledge Extraction for Taxonomies of Concepts from Wikipedia</i> .....	28
Kepa BENGOETXEA and Koldo GOJENOLA <i>Exploring Treebank Transformations in Dependency Parsing</i> .....	33
Walter BOSMA <i>Contextual Saliency in Query-based Summarization</i> .....	39
Aurélien BOSSARD and Thierry POIBEAU <i>Integrating Document Structure into a Multi-Document Summarizer</i> .....	45
Julian BROOKE, Milan TOFILOSKI and Maite TABOADA <i>Cross-Linguistic Sentiment Analysis: From English to Spanish</i> .....	50
Zdenek CESKA and Chris FOX <i>The Influence of Text Pre-processing on Plagiarism Detection</i> .....	55
Kostadin CHOLAKOV and Gertjan van NOORD <i>Combining Finite State and Corpus-based Techniques for Unknown Word Prediction</i> .....	60
Walter DAELEMANS, Hendrik J. GROENEWALD and Gerhard B. van HUYSSSTEEN <i>Prototype-based Active Learning for Lemmatization</i> .....	65
Heshaam FAILI <i>From Partial toward Full Parsing</i> .....	71
Ingrid FALK, Claire GARDENT, Evelyne JACQUEY and Fabienne VENANT <i>Grouping Synonyms by Definitions</i> .....	76
Francesca FALLUCCHI and Fabio Massimo ZANZOTTO <i>Singular Value Decomposition for Feature Selection in Taxonomy Learning</i> .....	82
Olivier FERRET <i>Improving Text Segmentation by Combining Endogenous and Exogenous Methods</i> .....	88

Kuzman GANCHEV and Georgi GEORGIEV <i>Edlin: an Easy to Read Linear Learning Framework</i> .....	94
Surya GANESH and Vasudeva VARMA <i>Exploiting the use of Prior Probabilities for Passage Retrieval in Question Answering</i> .....	99
Surya GANESH and Vasudeva VARMA <i>Exploiting Structure and Content of Wikipedia for Query Expansion in the Context of Question Answering</i> .....	103
Albert GATT and François PORTET <i>Text Content and Task Performance in the Evaluation of a Natural Language Generation System</i> .....	107
Georgi GEORGIEV, Preslav NAKOV, Kuzman GANCHEV, Petya OSENOVA and Kiril SIMOV <i>Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields</i> .....	113
Bénédicte GOUJON <i>Uncertainty Detection for Information Extraction</i> .....	118
Samer HASSAN and Rada MIHALCEA <i>Learning to Identify Educational Materials</i> .....	123
Hany HASSAN, Khalil SIMA'AN and Andy WAY <i>Lexicalized Semi-Incremental Dependency Parsing</i> .....	128
Martin HASSEL and Hercules DALIANIS <i>Identification of Parallel Text Pairs Using Fingerprints</i> .....	135
Christian Theil HAVE <i>Stochastic Definite Clause Grammars</i> .....	139
Leonhard HENNIG <i>Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis</i> .....	144
Yann Vigile HOAREAU, Adil EL-GHAL and Charles TIJUS <i>Detection of Opinions and Facts. A Cognitive Approach</i> .....	150
Arantza Díaz de ILARRAZA, Koldo GOJENOLA and Maite ORONoz <i>Evaluating the Impact of Morphosyntactic Ambiguity in Grammatical Error Detection</i> .....	155
Tomoya IWAKURA <i>Fast Boosting-based Part-of-Speech Tagging and Text Chunking with Efficient Rule Representation for Sequential Labeling</i> .....	161
Heng JI, Ralph GRISHMAN, Zheng CHEN and Prashant GUPTA <i>Cross-document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges</i> .....	166
Lidia KHYMLKO, Kilian A.FOTH and Wolfgang MENZEL <i>Co-Parsing with Competitive Models</i> .....	173
Manfred KLENNER, Stefanos PETRAKIS and Angela FAHRNI <i>Robust Compositional Polarity Classification</i> .....	180
Roman KLINGER and Christoph M. FRIEDRICH <i>Feature Subset Selection in Conditional Random Fields for Named Entity Recognition</i> .....	185

Roman KLINGER and Christoph M. FRIEDRICH <i>User's Choice of Precision and Recall in Named Entity Recognition</i> .....	192
Sandra KÜBLER and Desislava ZHEKOVA <i>Semi-Supervised Learning for Word Sense Disambiguation: Quality vs. Quantity</i> .....	197
Anna KUPŚĆ <i>Treelex Meets Adjectival Tables</i> .....	203
Egoitz LAPARRA and German RIGAU <i>Integrating WordNet and FrameNet using a Knowledge-based Word Sense Disambiguation Algorithm</i> ...	208
Adrien LARDILLEUX and Yves LEPAGE <i>Sampling-based Multilingual Alignment</i> .....	214
Hector LLORENS, Borja NAVARRO and Estela SAQUETE <i>Using Semantic Networks to Identify Temporal Expressions from Semantic Roles</i> .....	219
Diego Jesus de LUCENA and Ivandré PARABONI <i>The Design of an Experiment in Anaphora Resolution for Referring Expressions Generation</i> .....	225
Patrick McCRAE <i>A Model for the Cross-Modal Influence of Visual Context upon Language Processing</i> .....	230
Claudiu MIHĂILĂ and Dalila MEKHALDI <i>Bimodal Corpora Terminology Extraction: Another Brick in the Wall</i> .....	236
Tristan MILLER, Bertin KLEIN and Elisabeth WOLF <i>Exploiting Latent Semantic Relations in Highly Linked Hypertext for Information Retrieval in Wikis</i> .....	241
Petar MITANKIN, Stoyan MIHOV and Tinko TINCHEV <i>Large Vocabulary Continuous Speech Recognition for Bulgarian</i> .....	246
Emad MOHAMED and Sandra KÜBLER <i>Diacritization for Real-World Arabic Texts</i> .....	251
Karo MOILANEN and Stephen PULMAN <i>Multi-entity Sentiment Scoring</i> .....	258
Miguel A. MOLINERO, Benoît SAGOT and Lionel NICOLAS <i>A Morphological and Syntactic Wide-coverage Lexicon for Spanish: The Leffe</i> .....	264
Prakash MONDAL <i>How Limited is the Limit?</i> .....	270
Roser MORANTE, Vincent Van ASCH and Antal van den BOSCH <i>Dependency Parsing and Semantic Role Labeling as a Single Task</i> .....	275
Hajime MORITA, Hiroya TAKAMURA and Manabu OKUMURA <i>Structured Output Learning with Polynomial Kernel</i> .....	281
Claire MOUTON, Guillaume PITEL, Gaël de CHALENDAR and Anne VILNAT <i>Unsupervised Word Sense Induction from Multiple Semantic Spaces with Locality Sensitive Hashing</i> .....	287

Svetlin NAKOV, Preslav NAKOV and Elena PASKALEVA <i>Unsupervised Extraction of False Friends from Parallel Bi-Texts Using the Web as a Corpus</i> .....	292
Adeline NAZARENKO and Haïfa ZARGAYOUNA <i>Evaluating term extraction</i> .....	299
Matteo NEGRI and Milen KOUYLEKOV <i>Question Answering over Structured Data: an Entailment-Based Approach to Question Analysis</i> .....	305
Viet Cuong NGUYEN, Le Minh NGUYEN, and Akira SHIMAZU <i>A Semi-supervised Approach for Generating a Table-of-Contents</i> .....	312
Lionel NICOLAS, Miguel A. MOLINERO, Benoît SAGOT, Elena Sánchez TRIGO, Éric de La CLERGERIE, Miguel Alonso PARDO, Jacques FARRÉ and Joan Miquel VERGÉS <i>Towards efficient production of linguistic resources: the Victoria Project</i> .....	318
Rafael L. de OLIVEIRA, Eder M. de NOVAIS, Roberto P.A. de ARAUJÓ and Ivandré PARABONI <i>A Classification-driven Approach to Document Planning</i> .....	324
Daniel ORTIZ-MARTÍNEZ, Ismael GARCÍA-VAREA and Francisco CASACUBERTA <i>Interactive Machine Translation Based on Partial Statistical Phrase-based Alignments</i> .....	330
Michael PAUL and Roxana GIRJU <i>Topic Modeling of Research Fields: An Interdisciplinary Perspective</i> .....	337
Guy PERRIER <i>An Interaction Grammar of interrogative and relative clauses in French</i> .....	343
Marius POPESCU and Liviu P. DINU <i>Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis</i> .....	349
Taraka RAMA and Anil Kumar SINGH <i>From Bag of Languages to Family Trees From Noisy Corpus</i> .....	355
Veselin RAYCHEV and Preslav NAKOV <i>Language-Independent Sentiment Analysis Using Subjectivity and Positional Information</i> .....	360
Siva REDDY, Abhilash INUMELLA, Rajeev SANGAL and Soma PAUL <i>All Words Unsupervised Semantic Category Labeling for Hindi</i> .....	365
Vassiliki RENTOUMI, George GIANNAKOPOULOS, Vangelis KARKALETSIS and George A. VOUIROS <i>Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach</i> .....	370
Magnus ROSELL, Martin HASSEL and Viggo KANN <i>Global Evaluation of Random Indexing through Swedish Word Clustering Compared to the People's Dictionary of Synonyms</i> .....	376
Alla ROZOVSKAYA and Roxana GIRJU <i>Identifying Semantic Relations in Context: Near-misses and Overlaps</i> .....	381
Ricardo SÁNCHEZ-SÁEZ, Joan-Andreu SÁNCHEZ and José-Miguel BENEDÍ <i>Statistical Confidence Measures for Probabilistic Parsing</i> .....	388

Luís SARMENTO, Paula CARVALHO and Eugénio OLIVEIRA <i>Exploring the Vector Space Model for Finding Verb Synonyms in Portuguese</i> .....	392
Bálint SASS <i>A Unified Method for Extracting Simple and Multiword Verbs with Valence Information and Application for Hungarian</i> .....	398
Ravi SINHA and Rada MIHALCEA <i>Combining Lexical Resources for Contextual Synonym Expansion</i> .....	403
Jan ŠNAJDER and Bojana Dalbelo BAŠIĆ <i>String Distance-Based Stemming of the Highly Inflected Croatian Language</i> .....	410
Marina SOKOLOVA and Victoria BOBICEV <i>Classification of Emotion Words in Russian and Romanian Languages</i> .....	415
Marina SOKOLOVA and Guy LAPALME <i>Classification of Opinions with Non-affective Adverbs and Adjectives</i> .....	420
Martha Yifiru TACHBELIE and Wolfgang MENZEL <i>Amharic Part-of-Speech Tagger for Factored Language Modeling</i> .....	427
Vicent TAMARIT, Carlos-D.MARTÍNEZ-HINAREJOS and José-M. BENEDÍ RUÍZ <i>Improving Unsegmented Statistical Dialogue Act Labelling</i> .....	433
Sara TONELLI and Emanuele PIANTA <i>Three issues in cross-language frame information transfer</i> .....	440
Antonio TORAL, Óscar FERRÁNDEZ, Eneko AGIRRE and Rafael MUÑOZ <i>A Study on Linking Wikipedia Categories to Wordnet Synsets using Text Similarity</i> .....	448
Jantine TRAPMAN and Paola MONACHESI <i>Ontology engineering and knowledge extraction for crosslingual retrieval</i> .....	454
Jernej VIČIČ, Petr HOMOLA and Vladislav KUBOŇ <i>A Method to Restrict the Blow-up of Hypotheses of a Non-disambiguated Shallow Machine Translation System</i> .....	459
Marc VILAIN, Jonathan HUGGINS and Ben WELLNER <i>Sources of Performance in CRF Transfer Training: a Business Name-tagging Case Study</i> .....	464
Tong WANG and Graeme HIRST <i>Extracting Synonyms from Dictionary Definitions</i> .....	470
Holger WUNSCH, Sandra KÜBLER and Rachael CANTRELL <i>Instance Sampling Methods for Pronoun Resolution</i> .....	477
Torsten ZESCH and Iryna GUREVYCH <i>Approximate Matching for Evaluating Keyphrase Extraction</i> .....	483
Ziqi ZHANG, Lei XIA, Mark A. GREENWOOD and José IRIA <i>Too Many Mammals: Improving the Diversity of Automatically Recognized Terms</i> .....	489