Answer Validation through Textual Entailment

THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY (ENGINEERING)
OF
JADAVPUR UNIVERSITY

BY

PARTHA PAKRAY
Department of Computer Science & Engineering
Jadavpur University
Kolkata 700032

UNDER THE ESTEEMED GUIDANCE OF

PROF. (DR.) SIVAJI BANDYOPADHYAY

&

PROF. (DR.) ALEXANDER GELBUKH

May, 2013

1. Introduction (Chapter 1)

A **Question Answering (QA)** system is an automatic system capable of answering natural language questions in a human-like manner: with a short, accurate answer. A question answering system can be domain specific, which means that the topics of the questions are restricted. Often, this means simply that also the document collection, i.e., the corpus, in which the answer is searched, consists of texts discussing a specific field. This type of QA is easier, for the vocabulary is more predictable, and ontologies describing the domain are easier to construct. The other type of QA, open-domain question answering, deals with unrestricted topics. Hence, questions may concern any subject. The corpus may consist of unstructured or structured texts. Yet another way of classifying the field of QA deals with language. In monolingual QA both the questions and the corpus are in the same language. In cross-language QA the language of the questions (source language) is different from the language of the documents (target language). The question has to be translated in order to be able to perform the search. Multilingual systems deal with multiple target languages i.e., the corpus contains documents written in different languages. In multilingual QA, translation issues are thus central as well. The goal of question answering (QA) is to identify and present to the user an actual answer to a question, rather than identifying documents that may be topically related to the question or may contain the answer. During the last ten years or so, research in Mono-lingual/ Cross-lingual / Multi-lingual question answering systems has been vigorously pursued through the Cross-Language Evaluation Forum (CLEF)[1]. Question answering (QA) is an area that operates on top of search engines (Google, Yahoo, etc.) in order to provide users with more accurate and elaborated responses where search engine responses outputs remain punctual, difficult to understand, and sometimes incoherent.

Answer Validation deals with the development and evaluation of subsystems aimed at validating the correctness of the answers generated by a QA system. The Answer Validation Exercise (AVE)[2] task identifies whether an answer extracted from a document is a valid answer to the given question. The result can be either "VALIDATED" or "REJECTED", which means whether it is a valid answer or not. The automatic answer validation is useful for improving the QA system performance and helping humans in the assessment of QA systems output. Answer Validation Exercise (AVE) is a task introduced in the Question Answering track at the Conference and Labs of the Evaluation Forum (CLEF) competition. AVE task is aimed at developing systems that decide whether the answer of a Question Answering system is correct or not. First Answer Validation Exercise (Peñas et. al., 2006) task was proposed in the Question Answering Track at CLEF in 2006 to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by a QA system. The basic idea is that once a pair (answer and snippet) is returned by a QA system, a hypothesis is built by turning the pair (question and answer) into an affirmative form. If the related

[1] www.clef-campaign.org
[2] http://nlp.uned.es/clef-qa/ave/

text (a snippet or a document) semantically entails this hypothesis, then the answer is expected to be correct. This automatic Answer Validation (Peñas et. al., 2006) is useful for:

- Improving QA systems performance
- Helping humans in the assessment of QA systems output
- Improving systems self-score
- Dealing with closed questions and related inferences
- Machine Translation Evaluation
- Summary Evaluation

There were three AVE competitions: AVE in 2006 (Peñas et. al., 2006), AVE in 2007 (Peñas et. al., 2007) and AVE in 2008 (Rodrigo et. al., 2008) all organized by the Conference and Labs of the Evaluation Forum (CLEF). The Answer Validation Exercise Evaluation track has been discussed in Chapter 3.

Textual Entailment: Given two texts one is called "*Text*" denoted as T and another one is called "*Hypothesis*" denoted as H, the Textual Entailment task is to decide whether or not the meaning of H can be logically inferred from that of T.

Textual Entailment can be formally defined as:

✓ *A text T entails a hypothesis H, if H is true in every circumstance (possible world) in which T is true*.

✓ *A text T entails a hypothesis H if, typically, a human reading T would infer that H is most likely true*.

For example, the text T = "*John's assassin is in jail"* entails the hypothesis H = "*John is dead"*; indeed, if there exists one's assassin, then this person is dead. Similarly, T = "*Mary lives in France"* entails H = "*Mary lives in Europe"*. On the other hand, T = "*It was like hitting the jackpot"* does not entail H = *"It was like removing the lottery"*.

2. Textual Entailment Survey (Chapter 2)

Recognizing Textual Entailment (RTE) is one of recent challenges of Natural Language Processing (NLP). Textual Entailment is defined as a directional relationship between pairs of text expressions (Text and Hypothesis). Textual Entailment is a core NLP task and can be linked with several tasks in NLP such as:

i. ***Summarization (SUM)***, a summary should be entailed by the text.

ii. ***Paraphrases (PP)*** can be seen as mutual entailment between a text and a hypothesis.

iii. ***Information Extraction (IE)***, the extracted information should also be entailed by the text.

iv. ***Question Answering (QA)*** the answer obtained for one question after the Information Retrieval (IR) process must be entailed by the supporting snippet of text.

v. ***Machine Translation (MT)*** the system generated translation should be semantically equivalent to the gold standard translation, i.e., must entail each other.

Entailment can be defined as a relation that holds between two language expressions (i.e. a text T and a hypothesis H) if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T. The relation is directional as the meaning of one expression can entail the meaning of the other, but not true for the vice versa. The Recognizing Textual Entailment (RTE) Challenge (Dagan et. al., 2005) introduces a generic task that combines the semantic inferences required across NLP applications. Evolution of this task has been done through different RTE Challenges. Every challenge adds some new flavour to the task compared to its predecessors.

The first shared task of Recognizing Textual Entailment was proposed by Dagan et. al. (2005) which requires the participating systems to predict whether there exists a textual entailment relation between two given texts, usually denoted as text (T) and hypothesis (H).

There were various evaluation tracks organized since the year 2005. In every new competition several new features of Textual Entailment were introduced. Three Recognizing Textual Entailment (RTE) competitions[3] RTE-1 (Dagan et. al., 2005) in 2005, RTE-2 (Bar-Haim et. al., 2006) in 2006 and RTE-3 (Giampiccolo et. al., 2007) in 2007 were organized by *Pattern Analysis, Statistical Modelling and Computational Learning* (PASCAL) - the European Commission's IST-funded Network of Excellence for Multimodal Interfaces. In 2008, the fourth edition (RTE-4) (Giampiccolo et. al., 2008) of the challenge was organized by National Institute of Standards and Technology (NIST) in Text Analysis Conference[4] (TAC). The TAC RTE-5 (Bentivogli et. al., 2009) challenge in 2009 includes a separate search pilot along with the main task. The TAC RTE-6 challenge (Bentivogli et. al., 2010), in 2010, includes the Main Task and Novelty Detection Task along with RTE-6 KBP Validation Pilot Task. The TAC RTE-7 (Bentivogli et. al., 2011) in 2011, focused on recognizing textual entailment in two application settings: Summarization and Knowledge Base Population. In RTE-7, Main Task and Novelty Detection Task are based on Summarization setting and KBP Validation Task is based on Knowledge Base Population setting. The Evaluation of NLP and Speech Tools for Italian (EVALITA) in 2009[5] was an evaluation campaign of both Natural Language Processing and speech technologies for Italian language. The EVALITA Textual Entailment task (Bos et. al., 2009) includes the detection of inferential relationships between pairs of short texts in Italian Language. In 2010, Parser Training and Evaluation using Textual Entailment[6] was organized by SemEval-2. In 2011, Recognizing

[3] http://pascallin.ecs.soton.ac.uk/Challenges/RTE/
[4] http://www.nist.gov/tac/
[5] http://www.evalita.it/2009
[6] http://semeval2.fbk.eu/semeval2.php?location=tasks&area=%20Parsing

Inference in Text[7] (RITE) was introduced in National Institute of Informatics Test Collection for Information Retrieval System (NTCIR-9). Languages in NTCIR-9 RITE task were Japanese, Simplified Chinese and Traditional Chinese. NTCIR-9 RITE subtasks were Binary-class (BC) subtask and Multi-class (MC) subtask. After the successful first RITE, RITE-2 held in 2012 in NTCIR-10. Cross-Lingual Textual Entailment task (CLTE) (Negri et. al., 2012) organized by Evaluation Exercises on Semantic Evaluation (SemEval) in 2012 was introduced. The CLTE addresses textual entailment (TE) recognition under a new dimension (cross-linguality) and within a new challenging application scenario (content synchronization).

Textual Entailment task is a text classification problem. TE can be classified as two-way ("***ENTAILMENT***" or "***YES***"/ "***NO ENTAILMENT***" or "***NO***"), three-way ("***ENTAILMENT***" or "***YES***"/ "***NO ENTAILMENT***" or "***NO***"/ "***UNKNOWN***" or "***CONTRADICTORY***"), multi-way ("***FOWARD***"/"***BACKWARD***"/"***BI DIRECTIONAL***"/"***CONTRACDICTION***"/"***INDEPENDENT***"). The general architecture of Textual Entailment as a classification problem has been shown in Figure 1.

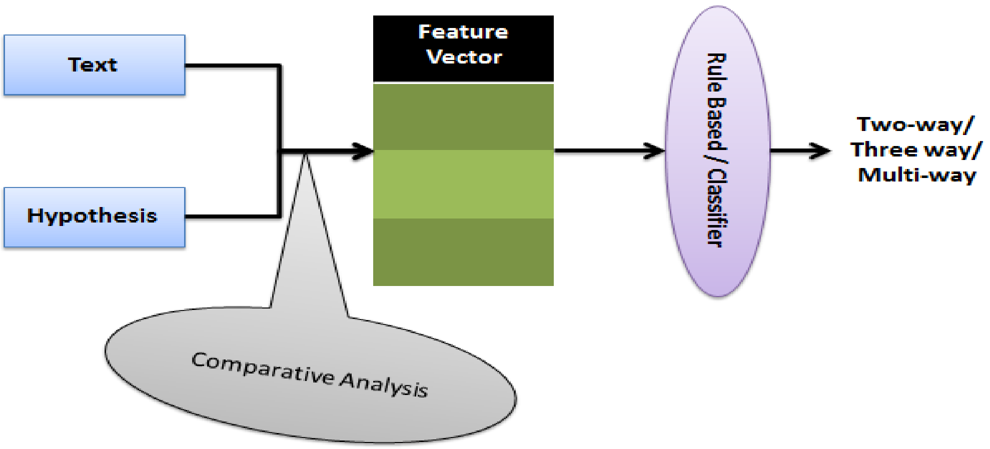

Figure 1: The General Architecture of Textual Entailment as a Classification Problem

Two-way RTE task was to decide whether:

i. ***T entails H*** **-** the pair would be marked as ENTAILMENT,

ii. ***T does not entail H*** **-** the pair would be marked as NO ENTAILMENT.

Three-way Classification was to decide whether:

i. ***T entails H*** **-** in which case the pair is marked as ***ENTAILMENT.***

ii. ***T contradicts H*** **-** in which case the pair is marked as ***CONTRADICTION.***

iii. ***The truth of H could not be determined on the basis of T*** **-** in which case the pair was marked as ***UNKNOWN.***

Multi-way Classification of Textual Entailment was to decide whether:

[7] http://research.nii.ac.jp/ntcir/ntcir-9/tasks.html

i. **Forward Entailment**: When the text infers the hypothesis but hypothesis does not infer the text.

ii. **Backward Entailment**: When the hypothesis infers the text but the reverse is not true.

iii. **Bidirectional Entailment**: When entailment occurs from both the direction, i.e., text infers hypothesis as well as hypothesis infers text.

iv. **Contradiction**: When text and hypothesis contradicts each other or cannot be true at the same time.

v. **Independent**: When forward, backward, bidirectional and contradiction relations does not hold.

An overview of Textual Entailment Evaluation Track from year 2005 to 2012 has been discussed. A large number of participants has participated in this track for experimenting their system and research on Textual Entailment domain. Lots of changes occurred from 2005 to 2012. Initially it was 2-way and 3-way classification problem after that it was converted into multiclass problem. There was also a paradigm shift from monolingual to cross-lingual systems. From the technical point of view, different kinds of techniques have been used to tackle the textual entailment problem by Lexical Level, Syntactic Level and also Semantic Level. Much effort is devoted by the NLP community to advances in Textual Entailment. Year to Year, interest of the field of textual entailment research has increased but the improvements in the textual entailment domain are still needed. Overall textual entailment survey has been discussed in Chapter 2.

3. Answer Validation Survey (Chapter 3)

In this chapter, Answer Validation Exercise (AVE) and the datasets, techniques and evaluation methods of Question Answering Evaluation tracks have been discussed. Answer Validation Exercise (AVE) is a task introduced in 2006 as part of the Question Answering (QA) evaluation track at Cross Lingual Evaluation Forum (CLEF) competition. AVE task is aimed at developing systems that validate the answer of a Question Answering system. There were three AVE competitions: AVE 2006 (Peñas et. al., 2006), AVE 2007 (Peñas et. al., 2007) and AVE 2008 (Rodrigo et. al., 2008). AVE systems receive a set of triplets (Question, Answer and Supporting Text) and return a judgment of “SELECTED”, “VALIDATED” or “REJECTED” for each triplet. The evaluation methodology was improved over the years and oriented to identify the useful factors for QA systems improvement. Thus, in 2007 (Peñas et. al., 2007) the AVE systems were to select only one “VALID” answer for every question from a set of possible answers, whereas in 2006 (Peñas et. al., 2006), several “VALID” answers were possible to be selected. In 2008 (Rodrigo et. al., 2008), the organizers increased the complexity of the data set by setting that all the answers to a question may be incorrect. The task of the participating systems was to ensure that all the answers to such questions are marked as “REJECTED”.

In Question Answering (QA) System one of the important tasks is that of correct answer selection from multiple choice questions. So in that case Answer Validation is one of the important tasks in QA system for answer ranking. An Answer Validation system can be applied in Question Answering systems to improve the ranking between possible answers. A Question Answering (QA) evaluation task over European legislation ResPubliQA (Peñas et. al., 2010) at CLEF 2010 has been described in this chapter. The Answer Validation Exercise evaluation track was renamed as Question Answering for Machine Reading Evaluation (QA4MRE) with more defined tasks. The Question Answering for Machine Reading Evaluation (QA4MRE) track was introduced at CLEF 2011 and the same has been been described in this chapter. Question Answering for Machine Reading Evaluation (QA4MRE) at CLEF 2012 has also been described in this chapter.

4. Textual Entailment System (Chapter 4)

In this chapter, several experiments that have been carried out for detecting the entailment between two texts have been reported. The experiments mainly focused on Lexical based Textual Entailment, Syntactic based textual entailment, Hybrid textual entailment, Machine Learning based textual entailment, Textual Entailment System with Embedded Anaphora Resolution, Semantic based Textual Entailment System, Multiclass Textual Entailment System and Cross Lingual Textual Entailment techniques. All textual entailment techniques have been discussed in Chapter 4.

The lexical Textual Entailment system (Pakray et. al., 2009) is based on the composition of the following six lexical RTE methods: WordNet based unigram match, Bigram match, Longest Common Sub-sequence, Skip-gram, Stemming and Named Entity matching. Each of these methods were trained on the development data to obtain two-way ("NO ENTAILMENT" or "No"/"ENTAILMENT" or "Yes") decisions. After that the system was applied on test data. The first experiment was carried out TAC RTE-5 datasets. This system participated in TAC RTE-5 Textual Entailment Evaluation Track. This experimented system has accuracy level of 58% for two-way decision.

Syntactic based Textual Entailment (Pakray et. al., 2010j) compares the dependency relations in both hypothesis and text. The system extracts syntactic structures from the text-hypothesis pairs using C&C CCG Parser[8] and Stanford Parser[9] separately and compares the corresponding structures to determine if the entailment relation is established. The system accepts pairs of text snippets (text and hypothesis) at the input and gives a value at the output: "YES" (or "ENTAILMENT") if the text entails the hypothesis and "NO" (or "NO ENTAILMENT") otherwise. The parser has been run and the dependency relations obtained for a text and hypothesis pair has been compared. Some of the important comparisons are: subject-subject comparison, subject-verb comparison, object-verb comparison and cross subject-verb comparison. Corresponding verbs are further compared using the

[8] http://svn.ask.it.usyd.edu.au/trac/candc/wiki
[9] http://nlp.stanford.edu/software/lex-parser.shtml

WordNet (Fellbaum, 1998). Each of the matches is assigned some weight learnt from the development corpus. A threshold has been set on the fraction of matching hypotheses relations based on the development set. The threshold score has been applied on the test set using the same methods of dependency parsing followed by comparisons. The TE system (Pakray et. al., 2010d) has participated in PETE (Yuret et. al., 2010) evaluation track; it has achieved 57% accuracy on PETE dataset.

The system based on Hybrid Textual Entailment (TE) recognition (Pakray et. al., 2011g) has used lexical and syntactic features. The important lexical similarity features that are used in the present system are: WordNet based uni-gram match, bi-gram match, longest common sub-sequence, skip-gram and stemming. In the syntactic TE system, the important features used are: subject-subject comparison, subject-verb comparison, object verb comparison and cross subject-verb comparison. This is the rule based system. The experiment has been carried out on RTE-4 test data and RTE-5 test data. It has achieved 55% on RTE-4 Data and 60% on RTE-5 test data. In Lexical based system, the score was 58% on RTE-5 dataset. It is clearly seen that performance of the hybrid TE system is improved by 2%.

Machine learning based Textual Entailment (TE) system (Pakray et. al., 2010a) that uses lexical similarity, lexical distance, chunk similarity and syntactic similarity features has been described in this chapter. The hybrid TE system is based on the Support Vector Machine and Naive Bayes classifier by WEKA[10] that uses the following features for training as well as the test set: Six features from Lexical TE, seventeen features from Lexical distance, one feature from POS similarity, one feature from chunk similarity and the eleven features from rule based syntactic two-way TE system. The system is trained on RTE datasets for building the model file. The system is tested on RTE-5 dataset and it has achieved 64% accuracy.

Semantic based Textual Entailment has been developed. This sematic based TE system (Pakray et. al., 2011f; Pakray et. al., 2011h) has used Universal Networking Language (UNL) relations (Uchida and Zhu, 2001). At first, the TE system identifies the UNL relations of both the Text and the Hypothesis by an En-Converter[11]. Then the TE system compares the UNL relations in both the text and the hypothesis to arrive at the two-way entailment decision. The system has been separately trained on each development corpus released as part of the Recognizing Textual Entailment (RTE) competitions RTE-1, RTE-2, RTE-3 and RTE-5 and tested on the respective RTE test sets.

The Cross-Lingual Textual Entailment is a new dimension (cross-linguality) of Textual Entailment domain and also a new challenging task for application scenario. The evaluation task of Cross-Lingual Textual Entailment (CLTE) Task was first introduced in SemEval 2012. The system (Neogi et. al., 2012a) for the CLTE task is based on a set of heuristics that assigns entailment scores to a text pair based on lexical relations. The text and the hypothesis in a text pair are translated to the same

[10] http//www.cs.waikato.ac.nz/~ml/weka
[11] http://unl.ru

language using the Microsoft Bing machine translation[12] system. The system separates the text pairs (T1 and T2) available in different languages and pre-processes them. After pre-processing the system uses several techniques for detecting textual entailment such as Word Overlaps, Named Entity matching, Chunk matching and POS matching to evaluate the separated text pairs. These modules return a set of score statistics, which helps the system to go for multi-class entailment decision based on the predefined rules.

The system (Pakray et. al., 2011c) for detecting Multiclass (MC) subtask is based on a learned system that uses different lexical similarity features like Word Net based Unigram Matching, Bigram Matching, Trigram Matching, Skip-gram Matching, Longest Common Subsequence (LCS) Matching and Named Entity (NE) Matching, Lexical Distance features and syntactic features. The system has calculated two entailment scores. One entailment score is S1 with T1 as Text and T2 as Hypothesis and the other entailment score is S2 with T1 as Hypothesis and T2 as Text. If the score S1, i.e., the mapping score with T1 as text and T2 as hypothesis is greater than the score S2, i.e., mapping score with T2 as text and T1 as hypothesis, then the entailment class will be "***forward***". Similarly if S1 is less than S2, i.e., T2 now acts as the text and T1 acts as the hypothesis then the entailment class will be "***backward***". Similarly if both the scores S1 and S2 are equal the entailment class will be "***bidirectional***" (entails in both directions). Measuring "bidirectional" entailment is much more difficult than any other entailment decision due to combinations of different lexical scores. As the system produces a final score (S1 and S2) that is basically the sum over different similarity measures, the tendency of identical S1 – S2 will be quite small. As a result system establishes another heuristic for "bidirectional" class. If the absolute value difference between S1 and S2 is below the threshold value, the system recognizes the pair as "bidirectional" *(abs (S1 – S2) < threshold)*. This threshold has been set as 5 based on observations on the training file. If the individual scores S1 and S2 fall below a certain threshold, again set based on the observations in the training file, the system concludes the entailment class as "***no-entailment***".

5. Answer Validation System (Chapter 5)

Answer validation deals with the development and evaluation of subsystems aimed at validating the correctness of the answers given by a Question Answering system. The automatic answer validation would be useful for improving Question Answering system performance, helping humans in the assessment of Question Answering systems output. An Answer Validation system receives the triplet Question, Candidate Answer and Supporting Text and returns a Boolean value indicating whether the Answer is correct for the Question according to the Supporting Text or not. The Textual Entailment techniques have been used to detect whether the entailment holds between the supporting text and the

[12] http://www.bing.com/translator

hypothesis generated from the question and the answer. The architecture of Answer Validation through Textual Entailment has been shown in Figure 2.

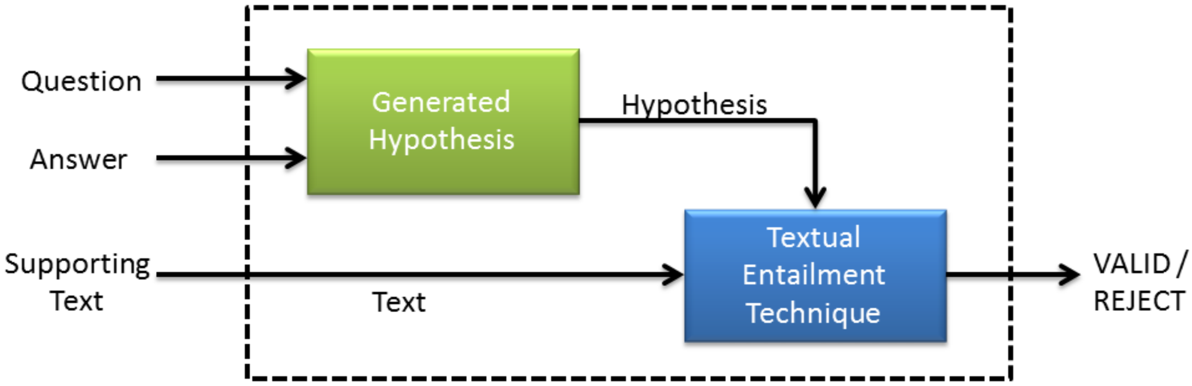

Figure 2: Architecture of Answer Validation system through Textual Entailment

The experimented systems identify the Answer Validation task into a Textual Entailment problem and use existing Textual Entailment system to validate the answers. So Textual Entailment is the main core component of Answer Validation system. Answer Validation plays an important part in Question Answering to detect whether the extracted answer is correct to the given question according to the relevant supporting text containing this answer. The supporting text of the answer forms the Text (**T**), and the question together with the answer forms the generated Hypothesis (**H**). An example is shown in Table 1. The Answer Validation Evaluation (AVE) task was defined as a RTE task in which the question and the answer constitute the hypothesis and the supporting text from which the answer was obtained is identified as the text.

Question	In what date was the first tennis championship at Wimbledon?
Supporting Text (i.e. T)	The first championships at Wimbledon, in London were played in 1877.
Answer	1877
Generated Hypothesis (i.e. H)	The first tennis championship at Wimbledon was in 1877.

Table 1: Example of Question Answering

If the entailment relation exists between the supporting text (i.e., T) and the hypothesis (i.e., H), the answer will be “VALIDATED”; otherwise, the answer may be wrong i.e. “REJECTED”.

In this thesis two experimented systems have been proposed for the Answer Validation task. One Answer Validation system is based on lexical and syntactic textual entailment and the other system is based one semantic textual entailment that uses the semantic features of Universal Networking Language.

The Answer Validation system (Pakray et. al., 2010f; Pakray, 2011; Pakray et. al., 2011e) is based on lexical and syntactic textual entailment. The important features used to develop the Answer

Validation system are Answer Pattern Generation, Hypothesis Generation, Textual Entailment, Named Entity Recognition (NER), Question-Answer Type Analysis and Chunk Boundary Module. The proposed Answer Validation (AV) system is rule based. The system first combines the question and the answer into Hypothesis (H) and the Supporting Text as Text (T) to identify the entailment relation as either “VALIDATED” or “REJECTED”. The system architecture has been shown in Figure 3.

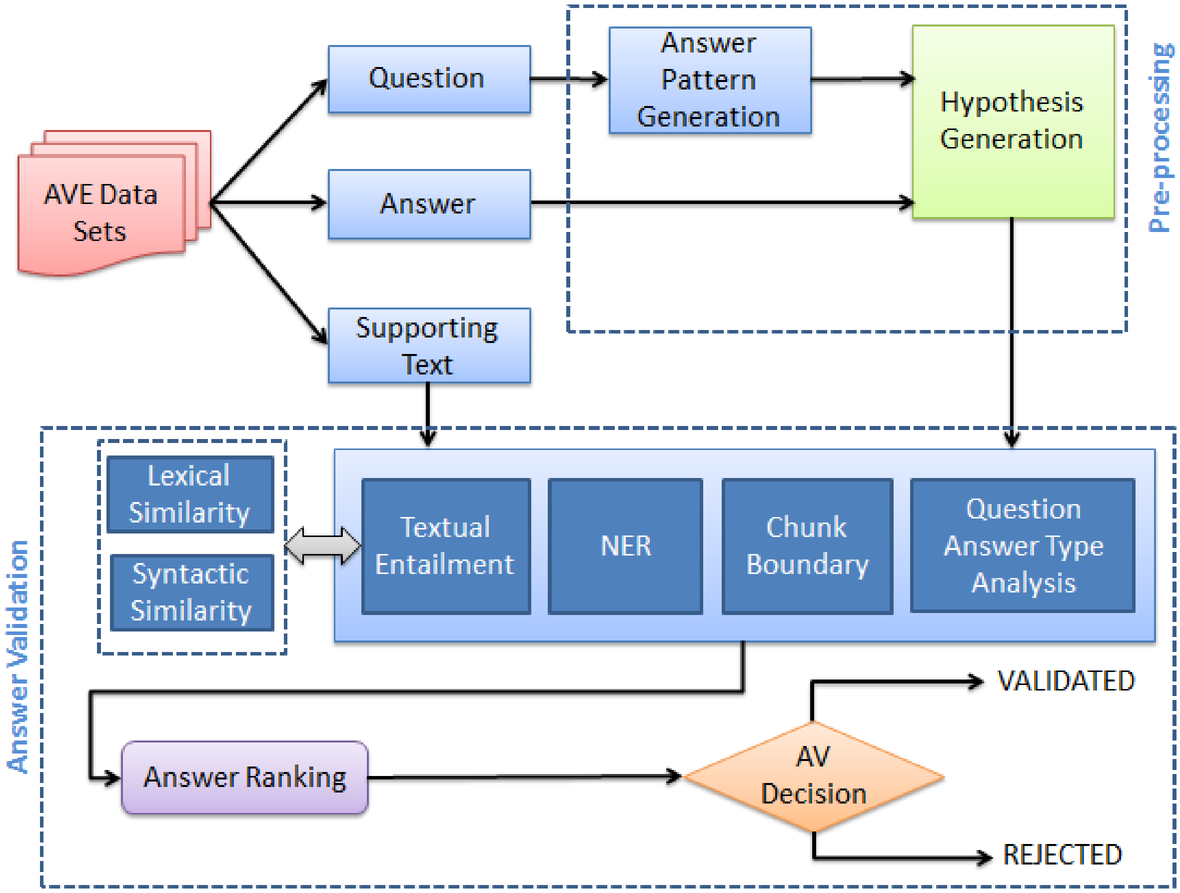

Figure 3: Architecture of Answer Validation System based on Lexical and Syntactic TE

Evaluation scores obtained on the AVE 2008 test set show 72% precision and 69% F-Score for “VALIDATED” decision.

The rule-based answer validation (AV) system (Pakray et. al., 2011d; Pakray et. al., 2012f) is based on Textual Entailment recognition mechanism that uses semantic features expressed in the Universal Networking Language (UNL) (Uchida and Zhu, 2001).

The following features of UNL are the motivating factors in its application in the Answer Validation system:

i. The most important information of a sentence is the **concept** that it conveys. This concept is represented by combination of UWs and Relations that are to be universal to the utmost, so that the representation is independent of language.

ii. **Time and Aspects**: This relates to the information on the tense and aspects in the sentence. These attributes are designed focusing on each part of point of an event of state that the sentence expresses.

iii. **Speaker's View of Reference and Communication**: This relates to the information related to the concept of a sentence. The attributes for expressing meanings of generality, specificity, negation, emphasis, focus, topic, etc. are included. Most of these attributes are designed focusing on each (part of) concept (s) in a sentence.

iv. **Speaker's Attitudes, Feelings and Viewpoints**: Information on the Mood Attributes is designed from a common perspective, although some of them cover special cases of some languages. In de-conversion or en-conversion of a native language, such attributes need to be transferred into or from its own expressions respectively.

v. **Ontological information** that is built-in to the UWs such as "icl" and "iof". For example singer(icl>musician>thing) which conveys the ontological information that a singer is a musician.

vi. **Logical information** such as "and" and "or" that are identified in the knowledge representation of a sentence in the UNL formalism.

vii. **Thematic information** such as "agt" = agent, "ins" = instrument, "tim" = time, "plc" = place, etc. that are identified in the knowledge representation of a sentence in the UNL formalism.

The architecture of the Answer Validation (AV) system (Pakray et. al., 2012f; Pakray et. al., 2011d) which is based on semantic textual entailment using Universal Networking Language (UNL) is presented in Figure 4. The main components of the AV system are: Pattern Generation Module, Hypothesis Generation Module, UNL en-conversion Module, Pre-processing module, Scoring Module and AVE Decision Module.

The experiments have been carried out on AVE 2008 Development Set and Test set for English. Evaluation scores obtained on the AVE 2008 test set show 69% precision and 65% F-Score for “VALIDATED” decision.

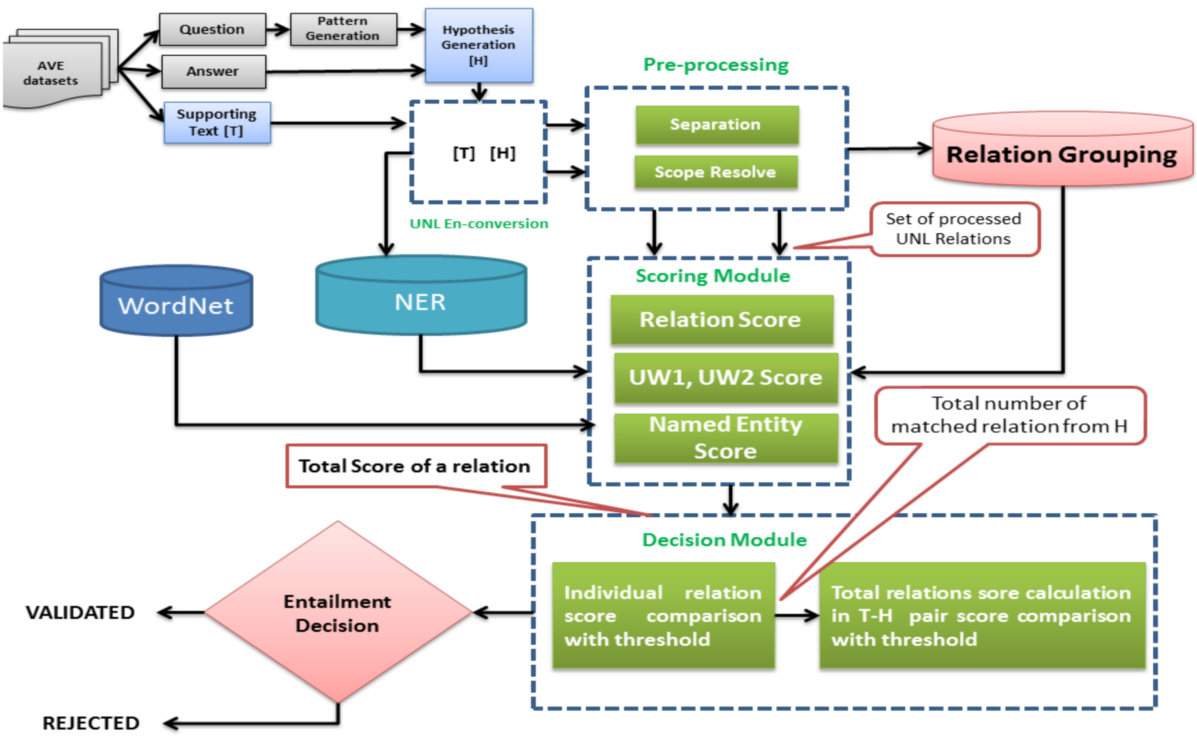

Figure 4: Architecture of Answer Validation system based on Semantic Textual Entailment using Universal Networking Language (UNL)

Question Answering (QA) is the one of the most challenging and demanding task in Natural Language Processing field. In QA Systems one of the important tasks is that of correct answer selection from multiple choice questions. So in that case Answer Validation is one of the important tasks in QA systems for answer ranking. An Answer Validation system can be applied in Question Answering system to improve the ranking between possible answers. The evaluation track for QA systems is organized by Cross-Language Evaluation Forum (CLEF) - RespubliQA[13] (Peñas et. al., 2010) in 2010, QA4MRE[14] (Peñas et. al., 2011) in 2011 and QA4MRE (Peñas et. al., 2012) in 2012. The main objective of QA4MRE is to develop a methodology for evaluating Machine Reading systems through Question Answering and Reading Comprehension Tests. Machine Reading task obtains an in-depth understanding of just one or a small number of texts. The task focuses on the reading of single documents and identification of the correct answer to a question from a set of possible answer options. Answer Validation technique (Pakray et. al., 2011e) has been applied to a Question Answering System (Bhaskar et. al., 2012) in QA4MRE at CLEF 2012. The experiment has been carried out on QA4MRE dataset. The architecture of the QA system based on Answer Validation has been shown in Figure 4.

[13] http://celct.isti.cnr.it/ResPubliQA/
[14] http://celct.fbk.eu/QA4MRE/

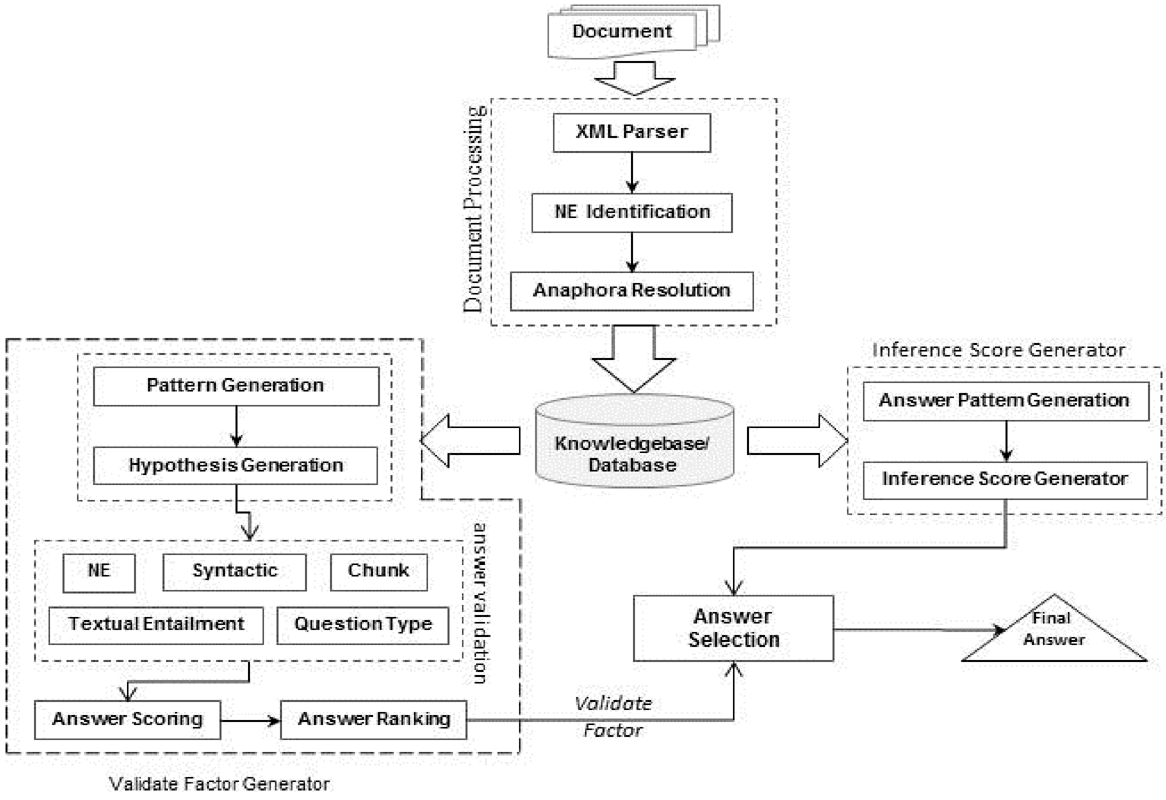

Figure 5: QA System based on Answer Validation

The developed system first combines the question and each answer option to form the Hypothesis, i.e., H. Stop words are removed from each H and query words are identified to retrieve the most relevant sentences from the associated document using Lucene. Relevant sentences are retrieved from the associated document based on the TF-IDF of the matching query words along with the n-gram overlap of the sentence with the H. Each retrieved sentence defines the Text, i.e., **T**. Each T-H pair is assigned a ranking score that works on textual entailment principle. A validate weight is automatically assigned to each answer options based on their ranking. A parallel procedure also generates the possible answer patterns from the given questions and answer options. Each sentence in the associated document is assigned an inference score with respect to each answer pattern. Evaluated inference score for each answer option is multiplied by the validate weight based on their ranking. The answer option that receives the highest selection score is identified as the most relevant option and is selected as the answer to the given question.

The first Answer Validation system (Pakray et. al., 2010f) has been applied in the QA4MRE system (Pakray et. al, 2011a) at CLEF 2011. The system (Pakray et. al, 2011a) has achieved 0.57 for c@1 evaluation measure in English Language. It was the best performing system among twelve participants. The system was improved in Textual Entailment module, answer pattern generation module and also change in scoring module. The improved system (Bhaskar et. al., 2012) participated in QA4MRE at CLEF 2012 Track. The QA4MRE@CLEF2012 system (Bhaskar et. al., 2012) has achieved accuracy of 0.65 for c@1 evaluation measure which is also the best performing system in eleven groups in English Language.

Another application of Answer Validation based on textual entailment is SMS-based FAQ retrieval (SMSFR) (Pakray et. al., 2012d). The system has used SMS-based FAQ retrieval dataset (Contractor et. al., 2011) of FIRE 2011[15] for this task. The goal of this task is to find a question Q from corpora of FAQs (Frequently Asked Questions) that best answers or matches the SMS query S. The system first checks the SMS using the Bing spell-checker. Then the system uses the Answer Validation module for monolingual FAQ retrieval.

6. Conclusion (Chapter 6)

An account of the key scientific contributions of this thesis along with a brief roadmap of the future possible avenues of this work has been reported in this chapter. The thesis makes key scientific research contributions in various areas of Textual Entailment, Answer Validation and Question Answering.

[15] http://www.isical.ac.in/~fire/faq-retrieval/

Publications on the Present Work

A. Journal Publications (3):

1. **Partha Pakray**, Utsab Barman, Sivaji Bandyopadhyay and Alexander Gelbukh. 2012(f). ***Semantic Answer Validation using Universal Networking Language***. In International Journal of Computer Science and Information Technologies (IJCSIT), ISSN 0975 - 9646, Pages 4927 - 4932, VOLUME 3 ISSUE 4 July- August 2012. http://www.ijcsit.com/docs/Volume%203/vol3Issue4/ijcsit2012030476.pdf

2. **Partha Pakray**, Soujanya Poria, Sivaji Bandyopadhyay, Alexander Gelbukh. 2011(h) ***Semantic Textual Entailment Recognition using UNL***. Polibits, ISSN 1870-9044, Issue 43, 2011, Pages 23–27.
http://polibits.gelbukh.com/2011_43/43-03.htm

3. **Partha Pakray**, Alexander Gelbukh and Sivaji Bandyopadhyay. 2011(g). ***TEXTUAL ENTAILMENT USING LEXICAL AND SYNTACTIC SIMILARITY***. In International Journal of Artificial Intelligence & Applications (IJAIA), Vol.2, No.1, January 2011, DOI : 10.5121/ijaia.2011.2104 , Pages 43-58.
http://airccse.org/journal/ijaia/papers/0111ijaia04.pdf

B. Book Chapters (5):

1. **Partha Pakray**, Snehasis Neogi, Sivaji Bandyopadhyay and Alexander Gelbukh. 2012(e). ***Recognizing Textual Entailment in Non-English Text via Automatic Translation into English***. In Proceedings of 11th Mexican International Conference on Artificial Intelligence, October 27 - November 4, 2012, San Luis Potosi, Mexico. Lecture Notes in Artificial Intelligence, Springer.
http://www.gelbukh.com/CV/Publications/2013/Recognizing%20Textual%20Entailment%20in%20Non-English%20Text%20via%20Automatic%20Translation%20into%20English.pdf

2. **Partha Pakray**, Snehasis Neogi, Sivaji Bandyopadhyay and Alexander Gelbukh. 2012(d) **SMSFR: *SMS-Based FAQ Retrieval System***. In Proceedings of 11th Mexican International Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence, N 7630, ISSN 0302-9743. October 27 - November 4, 2012, San Luis Potosi, Mexico.
http://www.gelbukh.com/CV/Publications/2013/SMSFR%20SMS-Based%20FAQ%20Retrieval%20System.pdf

3. **Partha Pakray**, Utsab Barman, Sivaji Bandyopadhyay and Alexander Gelbukh. 2011(f). ***A Statistics-Based Semantic Textual Entailment System***. 10th Mexican International Conference on Artificial Intelligence, November 26 - December 4, 2011, Puebla, Mexico. Lecture Notes in Artificial Intelligence N 7094, ISSN 0302-9743, Springer, 2011, p. 267-276.
   http://dx.doi.org/10.1007/978-3-642-25324-9_23

4. **Partha Pakray**, Alexander Gelbukh and Sivaji Bandyopadhyay. 2011(e). ***Answer Validation using Textual Entailment***. In proceedings 12th International Conference on Intelligent Text Processing and Computational Linguistics, February 20 to 26, 2011, Lecture Notes in Computer Science N 6609, ISSN 0302-9743, Springer, 2011, pp. 359-364.
   http://dx.doi.org/10.1007/978-3-642-22327-3_48

5. **Partha Pakray**, Alexander Gelbukh and Sivaji Bandyopadhyay. 2010(j). ***A Syntactic Textual Entailment System Using Dependency Parser***. Book Computational Linguistics and Intelligent Text Processing. Springer Berlin / Heidelberg, Volume Volume 6008/2010, ISBN 978-3-642-12115-9, Pages 269-278.
   http://dx.doi.org/10.1007/978-3-642-12116-6_22

C. Doctoral Symposium (3):

1. Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh Bandyopadhyay. 2013. ***Applications of Textual Entailment.*** In Doctoral Consortium 35th European Conference on Information Retrieval (ECIR) 2013. [Accepted].

2. Partha Pakray, Sivaji Bandyopadhyay and Alexander Gelbukh Bandyopadhyay. 2011(d). ***Answer Validation System through Textual Entailment and Universal Networking Language***. In Proceedings of the Doctoral Consortium at the 10th Mexican International Conference on Artificial Intelligence, MICAI-2011, November 28, 2011, Puebla, Mexico.ISBN 978-607-95367-4-9.
   http://dx.doi.org/10.1007/978-3-642-19437-5_29

3. Partha Pakray. 2011. ***Answer Validation using Textual Entailment***. In Proceedings of Natural Language Processing and Information Systems - 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Volume. 6716, ISBN 978-3-642-22326-6, pp. 324-329, Alicante, Spain, June28-30, 2011.
   http://dx.doi.org/10.1007/978-3-642-22327-3_48

D. Conference Publications (19):

1. **Partha Pakray**, Sivaji Bandyopadhyay and Alexander Gelbukh. 2013. ***Binary-class and Multi-class based Textual Entailment System***. NTCIR-10: The 10th NTCIR Workshop Meeting "Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access". RITE-2 competition: Recognizing Inference in TExt@NTCIR-10. National Institute of Informatics (NII), National Center of Sciences, June 18-21, 2013, NII, Tokyo, Japan.

2. **Partha Pakray**, Pinaki Bhaskar, Somnath Banerjee, Sivaji Bandyopadhyay and Alexander Gelbukh. 2012(c). ***An Automatic System for Modality and Negation Detection***. CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE). CLEF 2012 Labs and Workshop. Notebook Papers. 17-20 September 2012, Rome, Italy. http://www.clef-initiative.eu/documents/71612/a2e97aa6-4f75-4595-9a7a-da487053436c

3. Pinaki Bhaskar, **Partha Pakray**, Somnath Banerjee, Samadrita Banerjee, Sivaji Bandyopadhyay and Alexander Gelbukh. 2012. ***Question Answering System for QA4MRE@CLEF 2012.*** CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE). CLEF 2012 Labs and Workshop. Notebook Papers. 17-20 September 2012, Rome, Italy. http://www.clef-initiative.eu/documents/71612/d1bc614c-b465-4fbf-9e65-1b193dacd3ce

4. Snehasis Neogi, **Partha Pakray**, Sivaji Bandyopadhyay, Alexander Gelbukh. 2012(b) ***JU_CSE_NLP: Multi-grade Classification of Semantic Similarity between Text Pair.*** First Joint Conference on Lexical and Computational Semantics (*SEM), pages 571–574, Montreal, Canada, June 7-8, 2012. Association for Computational Linguistics. www.ixa2.si.ehu.es/starsem/proc/pdf/STARSEM-SEMEVAL083.pdf

5. Snehasis Neogi, **Partha Pakray**, Sivaji Bandyopadhyay and Alexander Gelbukh. 2012(a). ***JU_CSE_NLP: Language Independent Cross-lingual Textual Entailment System***. First Joint Conference on Lexical and Computational Semantics (*SEM), pages 689–695, Montreal, Canada, June 7-8, 2012. Association for Computational Linguistics. http://ixa2.si.ehu.es/starsem/proc/pdf/STARSEM-SEMEVAL103.pdf

6. **Partha Pakray**, Snehasis Neogi, Sivaji Bandyopadhyay, Alexander Gelbukh. 2011(c). ***A Textual Entailment System using Web based Machine Translation System***. NTCIR-9: The 9th NTCIR Workshop Meeting "Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access". RITE competition: Recognizing Inference in TExt@NTCIR9. National Institute of Informatics (NII), National Center of Sciences, Tokyo, Japan. December 6-9, 2011.
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/14-NTCIR9-RITE-PakrayP.pdf

7. **Partha Pakray**, Snehasis Neogi, Pinaki Bhaskar, Soujanya Poria, Sivaji Bandyopadhyay, Alexander Gelbukh. 2011(b). ***A Textual Entailment System using Anaphora Resolution***. Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook, November 14-15, 2011, National Institute of Standards and Technology, Gaithersburg, Maryland USA.
http://www.nist.gov/tac/publications/2011/participant.papers/JU_CSE_TAC.proceedings.pdf

8. **Partha Pakray**, Pinaki Bhaskar, Somnath Banerjee, Bidhan Chandra Pal, Sivaji Bandyopadhyay and Alexander Gelbukh. 2011(a). ***A Hybrid Question Answering System based on Information Retrieval and Answer Validation***. CLEF 2011 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE). CLEF 2011 Labs and Workshop. Notebook Papers. 19-22 September, Amsterdam. ISBN 978-88-904810-1-7; ISSN 2038-4726. 2011.
http://nlp.cic.ipn.mx/Publications/2011/Pakray_Clef2011.pdf

9. **Partha Pakray**, Santanu Pal, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander Gelbukh. 2010(i). ***JU_CSE_TAC: Textual Entailment Recognition System at TAC RTE-6***. System Report, Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook.
http://www.nist.gov/tac/publications/2010/participant.papers/JU_CSE_TAC.proceedings.pdf

10. **Partha Pakray**, Pinaki Bhaskar, Santanu Pal, Dipankar Das, Sivaji Bandyopadhyay and Alexander Gelbukh. 2010(g). ***JU_CSE_TE: System Description QA@CLEF 2010 – ResPubliQA.*** CLEF 2010 Workshop on Multiple Language Question Answering (MLQA 2010).
http://nlp.cic.ipn.mx/Publications/2010/JU_CSE_TE%20System%20Description.pdf

11. **Partha Pakray**, Santanu Pal, Sivaji Bandyopadhyay and Alexander Gelbukh. 2010(f) ***Automatic Answer Validation System on English Language.*** In IEEE 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE 2010), Chengdu, China, 20-22 August 2010, IEEE Catalog Number: CFP1052F-PRT, ISBN: 978-1-4244-6539-2, Volume 6, pp. 329-333.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5579166&isnumber=5579162

12. **Partha Pakray**, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2010(e). ***Textual Entailment and Anaphora Resolution***. In IEEE 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE 2010), Chengdu, China, 20-22 August 2010, IEEE Catalog Number: CFP1052F-PRT, ISBN: 978-1-4244-6539-2, Volume 6, pp. 334-336.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5579163&isnumber=5579162

13. **Partha Pakray**, Alexander Gelbukh and Sivaji Bandyopadhyay. 2010(d). ***Dependency Parser Based Textual Entailment System***. In IEEE 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI'10), IEEE CS, pp. 393-397.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5655646&isnumber=5654767

14. Santanu Pal, Tapabrata Mondal, **Partha Pakray**, Dipankar Das and Sivaji Bandyopadhyay. 2010(c). ***QGSTEC System Description – JUQGG: A Rule based approach***. In QG2010: The Third Workshop on Question Generation, The Tenth International Conference on Intelligent Tutoring Systems (ITS2010), Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 18, 2010.
http://oro.open.ac.uk/22343/1/QG2010-Proceedings.pdf

15. Santanu Pal, ***Partha Pakray***, Dipankar Das and Sivaji Bandyopadhyay. 2010(b). ***A Supervised Approach to Identify Semantic Relations from Paired Nominals***. In ACL-2010, SemEval 2010 Workshop, Uppsala, Sweden.
http://aclweb.org/anthology-new/S/S10/S10-1045.pdf

16. **Partha Pakray**, Alexander Gelbukh and Sivaji Bandyopadhyay. 2010(a). ***A Hybrid Textual Entailment System using Lexical and Syntactic Features***. In The 9th IEEE International Conference on Cognitive Informatics, (ICCI 2010), pp. 291-295.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5599726&isnumber=5599672

17. Pinaki Bhaskar, Amitava Das, **Partha Pakray**, Sivaji Bandyopadhyay. 2010. ***Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010.*** In FIRE 2010, Working Notes.
http://www.isical.ac.in/~fire/paper_2010/Pinaki-Fire-2010_SB_PB_AD_PP.pdf

18. **Partha Pakray**, Sivaji Bandyopadhyay, Alexander Gelbukh. 2009. ***Lexical based two-way RTE System at RTE-5.*** System Report, TAC RTE Notebook, 2009. http://www.nist.gov/tac/publications/2009/participant.papers/JU_CSE_TAC.proceedings.pdf

19. Srinivasa Rao Godavarthy, **Partha Pakray**, Sivaji Bandyopadhyay. 2007. ***Multilingual Restricted Domain QA System with Dialogue Management***. In proceedings of the Workshop on Cross Lingual Information Access, International Joint Conference on Artificial Intelligence (IJCAI 2007), 6-12 January 2007,Hyderabad, India, PP. 20-27.

E. Book

Partha Pakray. ***Multilingual Restricted Domain QA System with Dialogue Management***. LAP LAMBERT Academic Publishing (April 27, 2012), 96 pages, ISBN-10: 3845428112, ISBN-13: 978-3845428116.
http://www.amazon.com/Multilingual-restricted-domain-dialogue-management/dp/3845428112

References

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I. 2006. ***The Second PASCAL Recognising Textual Entailment Challenge***. Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy.
http://eprints.pascal-network.org/archive/00002675/01/RTE2-organizers.pdf

Bentivogli, L., Clark, P., Dagan, I., Dang, H. T., Giampiccolo, D. 2011. ***The Seventh PASCAL Recognizing Textual Entailment Challenge***. In TAC 2011 Notebook Proceedings.
http://www.nist.gov/tac/publications/2011/papers.html

Bentivogli, L., Dagan, I., Dang. H.T., Giampiccolo, D., Magnini, B. 2009. ***The Fifth PASCAL Recognizing Textual Entailment Challenge***. In TAC 2009 Workshop, National Institute of Standards and Technology Gaithersburg, Maryland USA.
http://www.nist.gov/tac/publications/2009/additional.papers/RTE5_overview.proceedings.pdf

Bentivogli, L., Magnini, B., Dagan, I., Dang, H.T., Giampiccolo, D. 2010. ***The Sixth PASCAL Recognizing Textual Entailment Challenge***. Proceedings of the Text Analysis Conference (TAC 2010) November 15-16, 2010 National Institute of Standards and Technology Gaithersburg, Maryland, USA.
http://www.nist.gov/tac/publications/2010/papers.html

Bos, J., Zanzotto, F. M., Pennacchiotti, M. 2009. ***Textual Entailment at EVALITA 2009***. EVALITA 2009 Procedings Textual Entailment Track.
http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2009/Textual%20Entailment/TE_ORGANIZERS.pdf

Contractor, D., Mittal, A., Padmanabhan D.S., Subramaniam L.V. 2011. ***SMS-based FAQ Retrieval***. FIRE 2011: Third Workshop of the Forum for Information Retrieval Evaluation, 2–4 December, IIT Bombay.
http://www.isical.ac.in/~fire/faq-retrieval/2011/faq-retrieval.html

Dagan, I., Glickman, O., Magnini, B. 2005. ***The PASCAL Recognising Textual Entailment Challenge***. Proceedings of the First PASCAL Recognizing Textual Entailment Workshop.
http://eprints.pascal-network.org/archive/00001298/01/dagan_et_al_rte05.pdf

Fellbaum, C. 1998. ***WordNet: An Electronic Lexical Database***. MIT Press, Cambridge, Mass.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., Cabrio, E. 2008. ***The Fourth PASCAL Recognizing Textual Entailment Challenge***. In TAC 2008 Proceedings.
http://www.nist.gov/tac/publications/2008/papers.html

Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B. 2007. ***The Third PASCAL Recognizing Textual Entailment Challenge.*** In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republi.
http://acl.ldc.upenn.edu/W/W07/W07-1401.pdf

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L.and Giampiccolo, D. 2012. ***Semeval-2012 Task 8: Crosslingual Textual Entailment for Content Synchronization***. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012).
http://ixa2.si.ehu.es/starsem/proc/pdf/STARSEM-SEMEVAL053.pdf

Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F. 2006. ***Overview of the Answer Validation Exercise 2006***. In Working Notes of CLEF 2006, Lecture Notes in Computer Science LNCS 4730. Springer, Berlin.
http://nlp.uned.es/docs/penasCLEF2006.pdf

Peñas, A., Forner, P., Rodrigo, A., Sutcliffe, R., Forascu, C. and Mota, C. ***Overview of ResPubliQA 2010: Question Answering Evaluation over European Legislation***. Working Notes of CLEF 2010.
http://celct.fbk.eu/ResPubliQA/scripts/downloadFile.php?file=/websites/ResPubliQA/resources/past_campaigns/2010/Documentations/ResPubliQA2010_Overview-final.pdf

Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Sutcliffe, R., Forascu, C., Sporleder, C. 2011. ***Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation***. In Working Notes of CLEF 2011.
http://celct.fbk.eu/ResPubliQA/scripts/downloadFile.php?file=/websites/ResPubliQA/resources/past_campaigns/2011/Documentations/QA4MRE-2011_Overview-final.pdf

Peñas, A., Hovy, E., Forner, P., Rodrigo, A., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P. 2012. ***Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation***. In Working Notes of CLEF 2012.
http://www.clef-initiative.eu/documents/71612/c076dd78-e36b-40d9-a6c8-fed4b7aa0b3d

Peñas, A., Rodrigo, Á., Verdejo, F. 2007. ***Overview of the answer validation exercise 2006***. In Working Notes of CLEF 2007.
http://nlp.uned.es/docs/penasCLEF2006.pdf

Rodrigo, A., Peñas, A. and Verdejo, F. 2008. ***Overview of the Answer Validation Exercise 2008***. In Working Notes of the CLEF 2008 Workshop. 17-19 September. Aarhus, Denmark. http://clef.isti.cnr.it/2008/working_notes/rodrigo-paperCLEF2008.pdf

Uchida, H., and Zhu, M. 2001. ***The Universal Networking Language beyond Machine Translation***. "International Symposium on Language in Cyberspace" held at 26 - 27 September 2001, Seoul of Korea, organized by The Korean National Commission for UNESCO and The Cyber Communication Academic Society in Korea, and sponsored by Institute of Information Technology Assessment, Rep. of Korea and UNESCO.
http://www.undl.org/publications/UNL-beyond%20MT.html

Yuret, D., Han, A., Turgut, Z. 2010. ***SemEval-2010 Task 12: Parser Evaluation using Textual Entailments***. Proceedings of the SemEval-2010 Evaluation Exercises on Semantic Evaluation. http://www.aclweb.org/anthology-new/S/S10/S10-1009.pdf