Vol-1172
urn:nbn:de:0074-1172-4

CLEF 2006
CLEF2006 Working Notes

Working Notes for CLEF 2006 Workshop
co-located with the 10th European Conference on Digital Libraries (ECDL 2006)

Alicante, Spain, September 20-22, 2006.

Edited by

Alessandro Nardi *
Carol Peters *
Jose Luis Vicedo **
Nicola Ferro (CEUR-WS Republishing Managing Editor) ***

* Istituto di Scienze e Tecnologie dell'Informazione (ISTI), CNR, Pisa, Italy
** Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Alicante, Spain
*** Department of Information Engineering (DEI), University of Padua, Via Gradenigo 6/B, 35131, Padova, Italy

Table of Contents

Preface

- What Happened in CLEF 2006: Introduction to the Working Notes
  Carol Peters
- Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them?
  Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro

Cross-Language Geographical Information Retrieval

- GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview
  Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio M. Di Nunzio, Nicola Ferro
- Report of MIRACLE Team for Geographical IR in CLEF 2006
  Sara Lana-Serrano, José M. Goñi-Menoyo, José C. González-Cristóbal
- GIR Experimentation
  Geoffrey Andogah
- Geographic IR Helped by Structured Geospatial Knowledge Resources
  A. Toral, O. Ferrández, E. Noguera, Z. Kozareva, A. Montoyo, R. Muñoz
- Monolingual and Bilingual Experiments in GeoCLEF2006
  Rocio Guillén
- University of Hagen at GeoCLEF2006: Experiments with Metonymy Recognition in Documents
  Johannes Leveling, Dirk Veiel
- UNSW at GeoCLEF 2006
  You-Heng Hu, Linlin Ge
- SINAI at GeoCLEF 2006: Expanding the Topics with Geographical Information and Thesaurus
  Manuel García-Vega, Miguel A. García-Cumbreras, L. Alfonso Ureña-López, José M. Perea-Ortega
- R2D2 at GeoCLEF 2006: a Mixed Approach
  Manuel García-Vega, Miguel A. García-Cumbreras, L. Alfonso Ureña-López, José M. Perea-Ortega, F. Javier Ariza-López, Oscar Ferrández, Antonio Toral, Zornitsa Kozareva, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, Davide Buscaldi, Paolo Rosso
- MSRA Columbus at GeoCLEF 2006
  Zhisheng Li, Chong Wang, Xing Xie, Wei-Ying Ma
- Place Disambiguation with Co-occurrence Models
  Simon Overell, João Magalhães, Stefan Rüger
- NICTA I2D2 Group at GeoCLEF 2006
  LiYi, Nicola Stokes, Lawrence Cavedon, Alistair Moffat
- Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006
  Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker
- WordNet-based Index Terms Expansion for Geographical Information Retrieval
  Davide Buscaldi, Paolo Rosso, Emilio Sanchis
- University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval
  Claudia Hauff, Dolf Trieschnigg, Henning Rode
- TALP at GeoCLEF-2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus

Daniel Ferrés, Horacio Rodríguez

- GeoCLEF Text Retrieval and Manual Expansion Approaches
  Ray R. Larson, Fredric C. Gey
- UB at GeoCLEF 2006
  Miguel E. Ruiz, Stuart Shapiro, June Abbas, Silvia B. Southwick, David Mark
- The University of Lisbon at GeoCLEF 2006
  Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade, Mário J. Silva

Cross-Language Retrieval in Image Collections

- Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks
  Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, Henning Müller
- Overview of the ImageCLEFmed 2006 Medical Retrieval and Annotation Tasks
  Henning Muller, Thomas Deselaers, Thomas Lehmann, Paul Clough, Eugene Kim, William Hersh
- Text Retrieval and Blind Feedback for the ImageCLEF Photo Task
  Ray R. Larson
- MIRACLE Team Report for ImageCLEF IR in CLEF 2006
  José Luis Martínez-Fernández, Julio Villena, Ana García-Serrano, Paloma Martínez
- Visual Micro-clustering Pre-processing for Cross-Language Ad hoc Image Retrieval
  Masashi Inoue
- Approaches Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval
  Yih-Chen Chang, Hsin-Hsi Chen
- Using Text and Image Retrieval Systems. Lic2m experiments at ImageCLEF 2006
  Romaric Besançon, Christophe Millet
- CELI Participation at ImageCLEF 2006: Comparison with the Ad-hoc Track
  Paolo Curtoni, Luca Dini, Vittorio Di Tomaso
- ImageCLEF 2006 Experiments at the Chemnitz Technical University
  Thomas Wilhelm, Maximilian Eibl
- IPAL Inter-Media Pseudo-Relevance Feedback Approach to ImageCLEF 2006 Photo Retrieval
  Nicolas Maillot, Jean-Pierre Chevallet, Vlad Valea, Joo Hwee Lim
- Dublin City University at CLEF 2006: Experiments for the ImageCLEF Photo Collection Standard Ad Hoc Task
  Kieran Mc Donald, Gareth J. F. Jones
- Performing Image Classification with a Frequency-based Information Retrieval Schema for ImageCLEF 2006
  Henning Muller, Tobias Gass, Antoine Geissbuhler
- University of Freiburg at ImageCLEF06 - Radiograph Annotation Using Local Relational Features
  Lokesh Setia, Alexandra Teynor, Alaa Halawani, Hans Burkhardt
- Morphosaurus in ImageCLEF 2006: The Effect of Subwords on Biomedical IR
  Philipp Daumke, Jan Paetzold, Kornél Markó
- Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006
  William Hersh, Jayashree Kalpathy-Cramer, Jeffery Jensen
- MedIC/CISMeF at ImageCLEF 2006: Image Annotation and Retrieval Tasks
  F. Florea, A. Rogozan, V. Cornea, A. Bensrhair, S. Darmoni
- Medical Image Annotation and Retrieval Using Visual Features
  Jing Liu, Yang Hu, Mingjing Li, Wei-ying Ma

- Combining Global Features within a Nearest Neighbor Classifier for Content-based Retrieval of Medical Images
  Mark O. Guld, Christian Thies, Benedikt Fischer, Thomas M. Lehmann
- Two-stage SVM for Medical Image Annotation
  Bo Qiu, Changsheng Xu, Qi Tian
- IPAL Knowledge-based Medical Image Retrieval in ImageCLEFmed 2006
  Caroline Lacoste, Jean-Pierre Chevallet, Joo-Hwee Lim, Xiong Wei, Daniel Raccoceanu, Diem Le Thi Hoang, Roxana Teodorescu, Nicolas Vuillenemot
- Baseline Results for the ImageCLEF 2006 Medical Automatic Annotation Task
  Mark O. Guld, Christian Thies, Benedikt Fischer, Thomas M. Lehmann
- Query and Document Translation by Automatic Text Categorization: A Simple Approach to Establish a Strong Textual Baseline for ImageCLEFmed 2006
  Julien Gobeill, Henning Müller, Patrick Ruch
- SINAI at ImageCLEF 2006
  M.C. Diaz-Galiano, M. A. García- Cumbreras, M. T. Martín-Valdivia, A. Montejo-Raez, L. A. Ureña-López
- CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections
  M. M. Rahman, Varun Sood, Bipin C. Desai, Prabir Bhattacharya
- Image Retrieval and Annotation Using Maximum Entropy
  Thomas Deselaers, Tobias Weyand, Hermann Ney

Cross-Language Speech Retrieval

- Overview of the CLEF-2006 Cross-Language Speech Retrieval Track
  Douglas W. Oard, Jianqiang Wang, Gareth J.F. Jones, Ryen W. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, Izhak Shafran
- The University of West Bohemia at CLEF 2006, the CL-SR Track
  Pavel Ircing, Ludek Muller
- Applying Logic Forms and Statistical Methods to CL-SR Performance
  R. M. Terol, P. Martínez-Barco, M. Palomar
- Speech Retrieval Experiments using XML Information Retrieval
  Djoerd Hiemstra, Roeland Ordelman, Robin Aly, Laurens van der Werff, Franciska de Jong
- University of Ottawa's Participation in the CL-SR Task at CLEF 2006
  Muath Alzghool, Diana Inkpen
- CLEF-2006 CL-SR at Maryland: English and Czech
  Jianqiang Wang, Douglas W.Oard
- Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments
  Gareth J. F. Jones, Ke Zhang, Adenike M. Lam-Adesina

Interactive Cross-Language Information Retrieval

- iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository
  Julio Gonzalo, Jussi Karlgren, Paul Clough
- Are Users Willing to Search Cross-Language? An Experiment with the Flickr Image Sharing Repository
  Javier Artiles, Julio Gonzalo, Fernando López-Ostenero,Víctor Peinado
- Providing Multilingual Access to FLICKR for Arabic Users
  Paul Clough, Azzah Al-Maskari, Kareem Darwish
- Trusting the Results in Crosslingual Keyword-based Image Retrieval
  Jussi Karlgren, Fredrik Olsson

Domain-Specific Information Retrieval

- Domain-Specific Track CLEF 2006: Overview of Results and Approaches, Remarks on the Assessment Analysis
  Maximilian Stempfhuber, Stefan Baerisch
- University of Hagen at CLEF2006: Reranking Documents for the Domain-specific Task
  Johannes Leveling
- Domain Specific Retrieval: Back to Basics
  Ray Larson
- Monolingual Retrieval Experiments with a Domain-Specific Document Corpus at the Chemnitz Technical University
  Jens Kursten, Maximilian Eibl

Multilingual Document Retrieval

- CLEF 2006: Ad Hoc Track Overview
  Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, Carol Peters
- CELI participation at CLEF 2006: Cross Language Delegated Search
  Paolo Curtoni, Luca Dini
- Oromo-English Information Retrieval Experiments at CLEF 2006
  Kula Kekeba Tune, Vasudeva Varma
- Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006
  Prasad Pingali, Vasudeva Varma
- Amharic-English Information Retrieval
  Atelach Alemu Argaw, Lars Asker
- Statistical Machine Translation and Cross-Language IR: QMUL at CLEF 2006
  Christof Monz
- The University of Lisbon at CLEF 2006 Ad-Hoc Task
  Nuno Cardoso, Mário J. Silva, Bruno Martins
- Evaluating Language Resources for English-Indonesian CLIR
  Herika Hayurani, Syandra Sari, Mirna Adriani
- Passage Retrieval vs. Document Retrieval in the Monolingual Task with the IR-n System
  Elisa Noguera, Fernando Llopis
- The PUCRS-PLN Group Participation at CLEF 2006
  Marco Gonzalez, Vera L. S de Lima
- Using Noun Phrases for Local Analysis in Automatic Query Expansion
  Joao Marcelo Azevedo Arcoverde, Maria das Graças Volpe Nunes, Wendel Scardua
- ENSM-SE at CLEF 2006: AdHoc Uses of Fuzzy Proximity Matching Function
  Annabelle Mercier, Michel Beigbeder
- A Study on the use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval
  Viviane Moreira Orengo
- Benefits of deep NLP-based Lemmatization for Information Retrieval
  Péter Halacsy
- Statistical vs. Rule-Based Stemming for Monolingual French Retrieval
  Prasenjit Majumder, Mandar Mitra, Kalyankumar Datta
- Report of MIRACLE Team for the Ad-Hoc Track in CLEF 2006
  José Miguel Goñi-Menoyo, José Carlos González-Cristóbal, Julio Villena-Román
- CoLesIR at CLEF 2006: Rapid Prototyping of an N-gram-Based CLIR System
  Jesús Vilares, Michael P. Oakes, John I. Tait
- SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: Query Expansion using the Google Search Engine

Fernando Martínez-Santiago, Artruro Montejo-Ráez, Miguel A. García-Cumbreras, L. Alfonso Ureña-López

- Robust Ad-hoc Retrieval Experiments with French and English at the University of Hildesheim
  Thomas Mandl, René Hackl, Christa Womser-Hacker
- Comparing the Robustness of Expansion Techniques and Retrieval Measures
  Stephen Tomlinson
- UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval
  Jacques Savoy, Samir Abdou
- REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval
  Angel Zazo, Carlos G. Figuerola, José Luis A. Berrocal
- DCU at CLEF 2006: Robust Cross Language Track
  Adenike M. Lam-Adesina, Gareth J.F. Jones

Multilingual Web Track

- Overview of WebCLEF 2006
  Krisztian Balog, Leif Azzopardi, Jaap Kamps, Maarten de Rijke
- REINA at WebCLEF2006. Mixing Fields to Improve Retrieval
  Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez, Emilio Rodríguez
- UPV/BUAP Participation in WebCLEF 2006
  David Pinto, Paolo Rosso, Ernesto Jiménez
- The University of Amsterdam at WebCLEF 2006
  Krisztian Balog, Maarten de Rijke
- Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for CLEF 2006 at the University of Hildesheim
  Ben Heuwing, Thomas Mandl, Robert Strötgen
- Text Reduction-Enrichment at WebCLEF
  Franco Rojas, Héctor Jiménez-Salazar, David Pinto
- European Web Retrieval Experiments at WebCLEF 2006
  Stephen Tomlinson
- Using Document Structure on Retrieving Webpages at the Web-CLEF 2006
  Syntia Wijaya, Bimo Widhi, Tommy Khoerniawan, Mirna Adriani

Multiple Language Question Answering

- Overview of the CLEF 2006 Multilingual Question Answering Track
  Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Petya Osenova, Anselmo Peñas, Valentin Jijkoun, Bogdan Sacaleanu, Paulo Rocha, Richard Sutcliffe
- Overview of the Answer Validation Exercise 2006
  Anselmo Penas, Álvaro Rodrigo, Valentín Sama, Felisa Verdejo
- Overview of WiQA 2006
  Valentin Jijkoun, Maarten de Rijke
- Experiments with LSA for Passage Re-Ranking in Question Answering
  David Tomas, José L. Vicedo, Empar Bisbal, Lidia Moreno
- Cross-Language French-English Question Answering using the DLT System at CLEF 2006
  Richard F. E. Sutcliffe, Kieran White, Darina Slattery, Igal Gabbay, Michael Mulcahy
- Prodicos Experiment Feedback for QA@CLEF2006
  Emmanuel Desmontils, Christine Jacquin, Laura Monceaux

- Cross-Lingual Question Answering by Answer Translation
  Johan Bos, Malvina Nissim
- Extraction of Definitions for Bulgarian
  Hristo Tanev
- Priberam's Question Answering System in a Cross-Language Environment
  Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, Cláudia Pinto, Daniel Vidal
- LCC's PowerAnswer at QA@CLEF 2006
  Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Jonathan Clark, Dan Moldovan
- The University of Groningen at QA@CLEF 2006: Using Syntactic Knowledge for QA
  Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, Jörg Tiedemann
- BRUJA System. The University of Jaén at the Spanish Task of CLEFQA 2006
  Miguel A. García-Cumbreras, L. Alfonso Ureña-López, Fernando Martínez-Santiago, Jose M. Perea-Ortega
- DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track
  Bogdan Sacaleanu, Günter Neumann
- Cross Lingual Question Answering using QRISTAL for CLEF 2006
  Dominique Laurent, Patrick Séguéla, Sophie Nègre
- CLEF2006 Question Answering Experiments at Tokyo Institute of Technology
  E. W. D. Whittaker, J. R. Novak, P. Chatain, P. R. Dixon, M. H. Heie, S. Furui
- The University of Amsterdam at CLEF@QA 2006
  Valentin Jijkoun, Joris van Rantwijk, David Ahn, Erik Tjong, Kim Sang, Maarten de Rijke
- Using a Text Summarization System for Monolingual Question Answering
  Pedro Paulo Balage Filho, Vinícius Rodrigues de Uzêda, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes
- The UPV at QA@CLEF 2006
  Davide Buscaldi, José Manuel Gomez, Paolo Rosso, Emilio Sanchis
- Developing a Question Answering System for the Romanian-English Track at CLEF 2006
  Georgiana Puscasu, Adrian Iftene, Ionut Pistol, Diana Trandabăt, Dan Tufiş, Alin Ceauşu, Dan Stefanescu, Radu Ion, Constantin Orăsan, Iustin Dornescu, Alex Moruz, Dan Cristea
- Finding Answers in the Œdipe System by Extracting and Applying Linguistic Patterns
  Romaric Besancon, Mehdi Embarek, Olivier Ferret
- Esfinge - a Modular Question Answering System for Portuguese
  Luís Costa
- INAOE at CLEF 2006: Experiments in Spanish Question Answering
  Antonio Juarez-Gonzalez, Alberto Téllez-Valero, Claudia Denicia-Carral, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda
- Accuracy of the AID System's Information Retrieval in Processing Huge Data Collections*
  Jolanta Mizera-Pietraszko
- A Shallow Approach for Answer Selection based on Dependency Trees and Term Density
  Manuel Perez-Coutino, Manuel Montes-y-Gómez, Aurelio López-López, Luis Villaseñor-Pineda, Aarón Pancardo-Rodríguez
- University of Hagen at QA@CLEF 2006: Interpretation and Normalization of Temporal Expressions
  Sven Hartrumpf, Johannes Leveling
- The LIA at QA@CLEF-2006
  Laurent Gillard, Laurianne Sitbon, Eric Blaudez, Patrice Bellot, Marc El-Bèze
- AliQAn and BRILI QA Systems at CLEF 2006
  S. Ferrandez, P. López-Moreno, S. Roger, A.. Ferrández, J. Peral, X. Alvarado, E. Noguera, F. Llopis

- The Bilingual System MUSCLEF at QA@CLEF 2006
  Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat, Michael Bagur, Kevin Séjourné
- MIRACLE at the Spanish CLEF@QA 2006 Track
  Cesar de Pablo-Sanchez, Ana González-Ledesma, Antonio Moreno, José Luis Martínez-Fernández, Paloma Martínez
- Finding Answers to Indonesian Questions from English Documents
  Sri Hartati Wijono, Indra Budi, Lily Fitria, Mirna Adriani
- Hunting Answers with RAPOSA (FOX)
  Luis Sarmento
- The Effect of Entity Recognition in Answer Validation
  Álvaro Rodrigo, Anselmo Penas, Felisa Verdejo
- A Knowledge-based Textual Entailment Approach applied to the QA Answer Validation at CLEF 2006
  O. Ferrandez, R.. M. Terol, R.. Muñoz, P. Martínez-Barco, M. Palomar
- Automatic Answer Validation using COGEX
  Marta Tatu, Brandon Iles, Dan Moldovan
- UNED Submission to AVE 2006
  Jesús Herrera, Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo
- Paraphrase Substitution for Recognizing Textual Entailment
  Wauter Bosma, Chris Callison-Burch
- Experimenting a “General Purpose” Textual Entailment Learner in AVE
  Fabio Massimo Zanzotto, Alessandro Moschitti
- University of Hagen at QA@CLEF 2006: Answer Validation Exercise
  Ingo Glockner
- Adaptation of a Machine-learning Textual Entailment System to a Multilingual Answer Validation Exercise
  Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo
- Towards Entailment-based Question Answering: ITC-irst at CLEF 2006
  Milen Kouylekov, Matteo Negri, Bernardo Magnini, Bonaventura Coppola
- The University of Amsterdam at WiQA 2006
  Sisay Fissaha Adafre, Valentin Jijkoun, Maarten de Rijke
- Identifying Novel Information using Latent Semantic Analysis in the WiQA Task at CLEF 2006
  Richard F. E. Sutcliffe, Josef Steinberger, Udo Kruschwitz, Mijail Alexandrov-Kabadjov, Massimo Poesio
- A Naïve Bag-of-Words Approach to Wikipedia QA
  Davide Buscaldi, Paolo Rosso
- University of Alicante at WiQA 2006
  Antonio Toral Ruiz, Georgiana Puşcaşu, Lorenza Moreno Monteagudo,Rubén Izquierdo Beviá, Estela Saquete Boró
- A High Precision Information Retrieval Method for WiQA
  Constantin Orasan, Georgiana Puşcaşu
- LexiClone Inc. at CLEF WiQA
  Ilya Geller
- MIRACLE at the Spanish WiQA Pilot: Using Named Entities and Cosine Similarity to extend Wikipedia articles
  Cesar de Pablo-Sanchez, José Luis Martínez-Fernández, Paloma Martínez

Appendix

- Appendix A: Results of the Ad-hoc Bilingual and Monolingual Tasks

- Appendix B: Results of the Ad-hoc Robust Task
- Appendix C: Results of the Domain-Specific Track
- Appendix D: Results of the GeoCLEF Track
- Appendix E: Statistics for the QA@CLEF Track
- Appendix F: Participating Institutions

2014-08-12: submitted by Nicola Ferro
2014-08-17: published on CEUR-WS.org |valid HTML5|

Finding Answers to Indonesian Questions from English Documents

Sri Hartati Wijono, Indra Budi, Lily Fitria, and Mirna Adriani

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
{shw50, indra, lifi50, mirna}@cs.ui.ac.id

Abstract. Our report describes the results of work in our participation in the Indonesian-English question-answering task of the 2006 Cross-Language Evaluation Forum (CLEF). In this work we translated an Indonesian query set into English using a machine translation tool available on the internet. Documents relevant to a question are first retrieved. The relevant documents are then divided into passages of 5 sentences each. The answer to the question is extracted from a passage relevant to the query. The answer is identified based on matching annotations between the query and the documents. A linguistic tool is used to annotate the words in the queries and the documents.

1 Introduction

This year we participate in the bilingual Question-Answering (QA) task, the Indonesian-English QA, of the Cross Language Evaluation Forum (CLEF) 2006. Finding the correct answer to a question in documents is a challenging task, and this is the main research topic in CLEF Question Answering task. The question and the documents must be analyzed in order to find the right answer in the documents. There are several techniques that have been used to handle the QA task, i.e., parsing and tagging the sentence in the question [6], in the documents [4], in paragraphs [8], and in passages [2, 3].

2 The Question Answering Process

The process of finding the answer to a query in documents proceeds in a number of stages. First, the original CLEF English queries were translated into Indonesian manually. Next, we classified the Indonesian questions (queries) according to the type of question. We identified the question type from the question word used in the query. The Indonesian question was then translated back into English using a machine translation tool. We used a web machine translation tool called *Toggletext*[1] to translate an Indonesian query set into English. We learned from our previous work [1] that freely available dictionaries on the Internet did not provide sufficiently good translation terms, as their vocabulary was very limited. We hoped that we could achieve better results using a machine translation approach.

The resulting English query was then used to retrieve the relevant documents from the collection by means of an information retrieval system. The contents of a number of documents at the top of the rank list were then split into passages. The passages were then tagged using a linguistic tagging (annotation) tool to identify the type of words in the passages. Finally, the passages were then scored using an algorithm, and the answer to the question is extracted from the passage with the highest score.

2.1 Categorizing the Questions

Each question category, which is identified by the question word in the question, points to the type of answer that is looked for in the documents. The Indonesian question-words used in the categorization are:

dimana, dimanakah, manakah (where)	points to <location>
apakah nama (what),	points to <location>

[1] See "http://www.toggletext.com".

siapa, siapakah (who)	points to <person>
berapa (how many)	points to <measure>
kapan (when)	points to <date>
organisasi apakah (what organization)	points to <organization>
apakah nama (which)	points to <location>
sebutkan (name)	points to < other>

By identifying the question type, we can predict the kind of answer that we need to look for in the document. The Indonesian question was tagged using a question tagger that we developed according to the question word that appears in the question. This approach is similar to those used by Clark et al. [2] and Hull [4,7,8]. However, we ignored the tagging on the question when we ran the query through the IR system to retrieve the documents.

2.2 Building Passages

The Indonesian question was translated into English using machine translation. The resulting English query was then run through an information retrieval system as a query to retrieve a list of relevant documents. We used *Lemur*[2] information retrieval system to index and retrieve the documents. The contents of the top 10 relevant documents were split into passages. Each passage contains five sentences where the last sentence is repeated in the next passage as the first sentence. The sentence in the documents was identified using a sentence parser to identify the beginning and the end of a sentence. The passages are then indexed by Lemur and the queries were run through to get the top-10 passages. These top-10 passages are being scored in order to get the answers to the queries.

2.3 Tagging the Passage

The passages were then run through an entity tagger to get the entity annotation tags. The entity annotation tagger identifies words of known entity types, and tags them with the entity type tags, such as person, location, and organization. For example, <*organization*> *UN*, the word *UN* is identified as an organization so it gets the organization tag. In this work, we used linguistic tagger tool, *Gate*[3].

Gate analyzes English words and annotates them with tags to indicate location, organization, and person, where applicable. The annotation tags are used to find the candidate answer based on the type of the question, for example, a word with location tag is a good candidate answer to a *where* question, and a word with a person tag is a good candidate answer to a *who* question.

2.4 Scoring the Passages

Passages were scored based on their probability of answering the question. The scoring rules are as follows:

1. Give 1 to a passage if its tag is not the same as the query tag and 0 if not.
2. Add 1 if a word in the passage is the same as the query.
3. Add 1 if the number of words in the passage is more than half of the number of the query words.

Once the passages obtained their scores, the top 10 scoring with the appropriate tags – e.g., if the question type is person (the question word "*who*") then the passages must contain the person tag – were then taken to the next stage.

2.5 Finding the Answer

The top 10 passages were analyzed to find the best answer. The probability of a word being the answer to the question is inversely proportional to the number of words in the passage that separate the candidate word and the word in the query. For each word that has the appropriate tag, its distance from a query word found in the passage is computed. The candidate word that has the smallest distance is the final answer to the question.

---

[2] See "http://www.lemurproject.org/".
[3] See "http://www.gate.shef.ac.uk/".

For example:

- Question: What is the **capital** of <*LOCATION*> Somalia?
- Passage:
  - Here there is no coordination. <*PERSON*> Steffan de Mistura – UNICEF representative in the Somali **capital**, <*LOCATION*> **Mogadishu**, and head of the anti-cholera team – said far more refugees are crowded together here without proper housing or sanitarian than during the <*LOCATION*> **Somalia** crisis. And many are already sick and exhausted by the long trek from <*LOCATION*> **Rwanda**.

The distance between the question word *capital* and *Mogadishu* is 1, between the question word *capital* and *Rwanda* is 38. So, *Mogadishu* becomes the final answer since its distance to the question word *capital* is the smallest one (closest).

3 Experiment

Our work focused on the bilingual task using Indonesian questions to find answers in English documents. The Indonesian questions were obtained by manually translating the English questions. The Indonesian questions were then translated back into English an online machine translation tool *Toggletext* to retrieve relevant English documents from the collection.

Using the *Gate* tagger to tag words in the passages, only 14 correct answers were found (see Table 1). There were 4 inexact (ambiguous) answers and 159 wrong answers.

Table 1. Evaluation of the QA result

Task : Bilingual QA	Evaluation
W (wrong)	159
U (unsupported)	13
X (inexact)	4
R (right)	14

Our result shows that we need to find ways to improve the effectiveness in finding correct answers, in particular, ways of reducing the number incorrect word tagging. We also learned that expanding the translated queries by adding related terms could also help, as more relevant documents can be retrieved for the QA algorithm to work.

4 Summary

Our participation in the QA task still needs further improvement. In our recent work, we managed to improve the QA result by applying query expansion and different passage scoring techniques. We hope that applying such techniques will result in a better performance next year.

References

1. Adriani, M. and van Rijsbergen, C. J. Term Similarity Based Query Expansion for Cross Language Information Retrieval. In Proceedings of Research and Advanced Technology for Digital Libraries (ECDL'99). Springer Verlag, Paris (1999) 311-322
2. Clarke, C. L. A., Cormack, G.G., Kisman, D. I. E. and Lynam, K. Question Answering by Passage Selection. In NIST Special Publication: The 9th Text retrieval Conference (2000)
3. Clarke, Charles L.A., Cormack, Gordon V. and Lynam, Thomas R. Exploiting Redundancy in Question Answering. In Proceeding of ACM SIGIR. New Orleans (2001)
4. Hull, David. Xerox TREC-8 Question Answering Track Report. In NIST Special Publication: The 8th Text Retrieval Conference (1999)
5. Li, Xiaoyan dan Croft, Bruce. Evaluating Question-Answering Techniques in Chinese. In NIST Special Publication: The 10th Text Retrieval Conference (2001)
6. Manning, C.D. and Schutze, H. Foundations of Statistical Natural Language Processing. The MIT Press, Boston (1999)

7. Moldovan, D. et.al. Lasso: A Tool for Surfing the Answer Net. In NIST Special Publication: The 8th Text Retrieval Conference (1999)
8. Pasca, Marius and Harabagiu, Sanda. High Performance Question Anwering. In Proceeding of ACM SIGIR. New Orleans (2001)