Email Ontology Learning System Based on Fuzzy Logic

Majdi Beseiso[1], Abdul Rahim Ahmad[2], Roslan Ismail[2]
[1]Yanbu University College
Yanbu, Saudi Arabia
[2]Universiti Tenaga Nasional (UNITEN)
Selangor, Malaysia
majdibsaiso@yahoo.com, {abdrahim, roslan}@uniten.edu.my

ABSTRACT: *Natural languages all over the world are context- based. There is inherent vagueness present in the common usage of language, but it is easily understood by people because they can decipher the context of the communication easily. Emails are an extension of face-to-face communication and many elements of ambiguity are present in the language used within the body of emails. In a computer based application to extract the ontologies for multilingual emails, it is important to disambiguate the sense of the verb, noun and adjective in a sentence develop correct hierarchies of relations and concepts. We propose to use a fuzzy logic based system as it is quite suitable for identifying and addressing the vagueness and ambiguity in the natural languages in the context of email ontologies. For various domains of emails like 'Meeting & Schedules', 'Comments &Discussions' etc., we have developed fuzzy rules and corresponding associations. These are used to develop an efficient learning model for email ontologies and incorporate in the JAPE pattern rules for extraction of appropriate and context aware concepts and relations. Implementation of fuzzy logic improves the efficiency and simplicity of the design process. In this research work we have tried to tackle the ambiguity of pronouns which act as important agents in ontology development. This results in accurate representation of ontologies in various domains of emails.*

Keywords: Fuzzy Logic, Ontology Learning, Coreference Resolution, Semantic Web

Received: 1 December 2012, Revised 18 January 2013, Accepted 21 January 2013

1. Introduction

With the advent of the internet and related technologies, the globalization and interaction among different cultures and countries is increasing at an exponential rate. The organizations and large corporations today are located in various countries employing and interacting with people of diverse cultural backgrounds and linguistic preferences. Email communication is an integral ingredient of every organization and a great source of knowledge within the organization. An effective and efficient knowledge management of Emails is critical for organizational growth. In this research, we propose to extend the research on Semantic Web technologies into the realm of Emails and especially multilingual Emails. The aim is to propose and build a semantic Email application capable of filtering, categorizing and extracting important information from emails automatically without human intervention. An important step in this direction is to semantically annotate the emails and generate ontologies for multilingual emails. With the development of the e-mail services and the clients, there are improvements of handling large email archives and offering search and tagging functionalities. However, these functionalities are still not mature, especially when the e-mail archive is large, not well organized and the tasks are related to multi-users and subtasks. In this case, most of the time, tedious

manual work is still inevitable. To sum up, there are three major problems with the current email systems:

1. Lack of semantic descriptions of email contents, so the email archiving and retrieval is difficult.

2. Poor integration among current email systems and other related systems, such as Microsoft office products in a semantic way.

3. Lack of multi-language support for semantic emails, which may, for example, cause difficulties in global business environment.

In the development of ontology learning algorithm to tackle above issues for emails, one peculiar problem is the informal nature of language used in email communication. For human beings, communication is contextual and the hidden meanings are intuitively understood based on the context of the discussion. Natural languages, the world over, have evolved over several thousands of years and components of ambiguity and vagueness are inherent in each language. These ambiguities are easily understood by human beings and their correct meanings are interpreted. But for a computer based system or an algorithm, such ambiguities and vagueness can result in erroneous results if not properly handled.

Ontology learning and semantic tagging for emails requires building up hierarchies of concepts and relations for different domains. These concepts and relations must be properly defined to capture the domain ontologies correctly. Any ambiguity or vagueness needs to be resolved before generating the RDF dataset for the ontologies. This research explores the usage of "*Fuzzy Logic*" in disambiguation of word senses so as to generate correct ontologies for emails in different domains. especially for coreference resolution where it is an important problem in domain of emails. Most work done on coreference resolution has focused on the news corpora and it is unclear how the performance would be in other real world domains. Also, the previous proposed methods are not tested for different languages, though a lack of corpora currently hinder research in this area [1].

2. Domain Ontologies for Email

Four basic domain vocabularies are considered and defined for the emails: News, Discussion and Comments, Meetings & Schedule, and Collaborative and Technical Requests in our study. Figure 1 demonstrates the use of LODE vocabulary for "*Schedule and Meetings*" domain. An "Event" is a central concept in the LODE vocabulary which can be used to relate "*ec: MeetingEmail*" to "*Event*" by "*ec: hasEvent*" property. Various LODE properties can be used to relate various terms in e-mail related to "*Schedule and Meetings*" domain like "*atPlace*", "*atTime*" and "*involdedAgent*" to get the time & duration, location and persons required for the meeting.

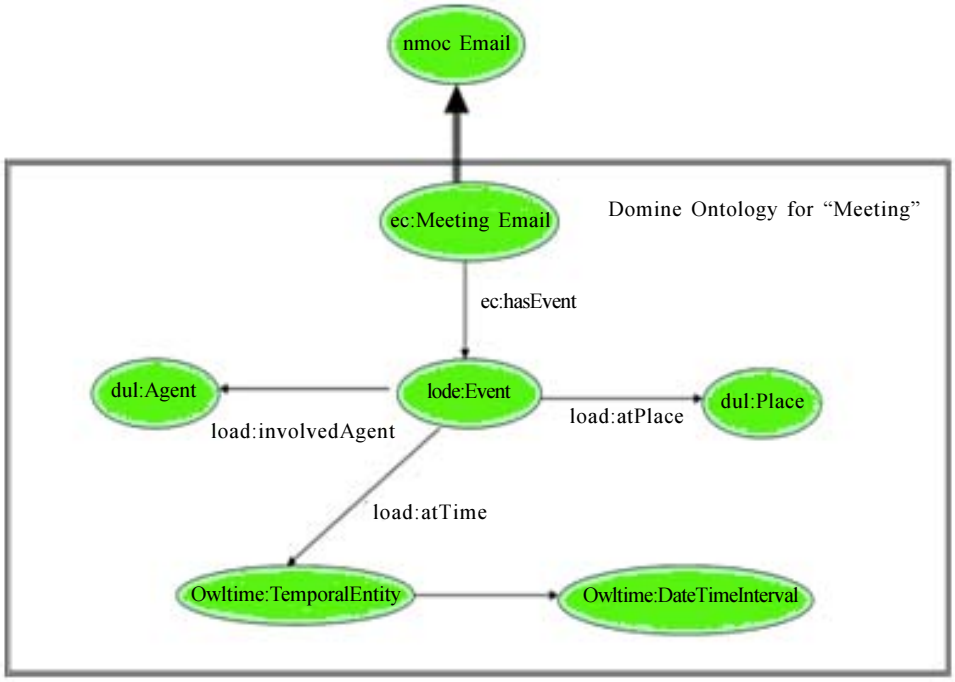

Figure 1. Domain Ontology for Email "*Meeting & Schedule*"

Similarly we have basic vocabularies for other domains and our task is to determine these relations and nodes for the four domains of emails mentioned above. For "*News*" domain, there are many existing ontologies. The most widely used one is the rNews which was developed by the International Press Telecommunications Council (IPTC). NewsItem is the central concept in the rNews ontology and it is related to the body of news content, the date and place of publishing, the multimedia objects, the user comments, etc. So we can relate "*ec:NewsEmail*" class to "*NewsItem*" by "*ec:hasNewsitem*", "*ec: hasArticle*" and *"ec: hasOrganization*" properties. "*NewsItem*" class can further be related to "*dateCreated*", "*headline*" and "*Agent*" properties as shown in Figure 2 below, which represents the vocabulary for news domain specific to emails. "*Organization*" will have associated properties as "*CountryName*", "*tel* " and "*Locality*". "*Article*" has a relation defined by "*hasSource*".

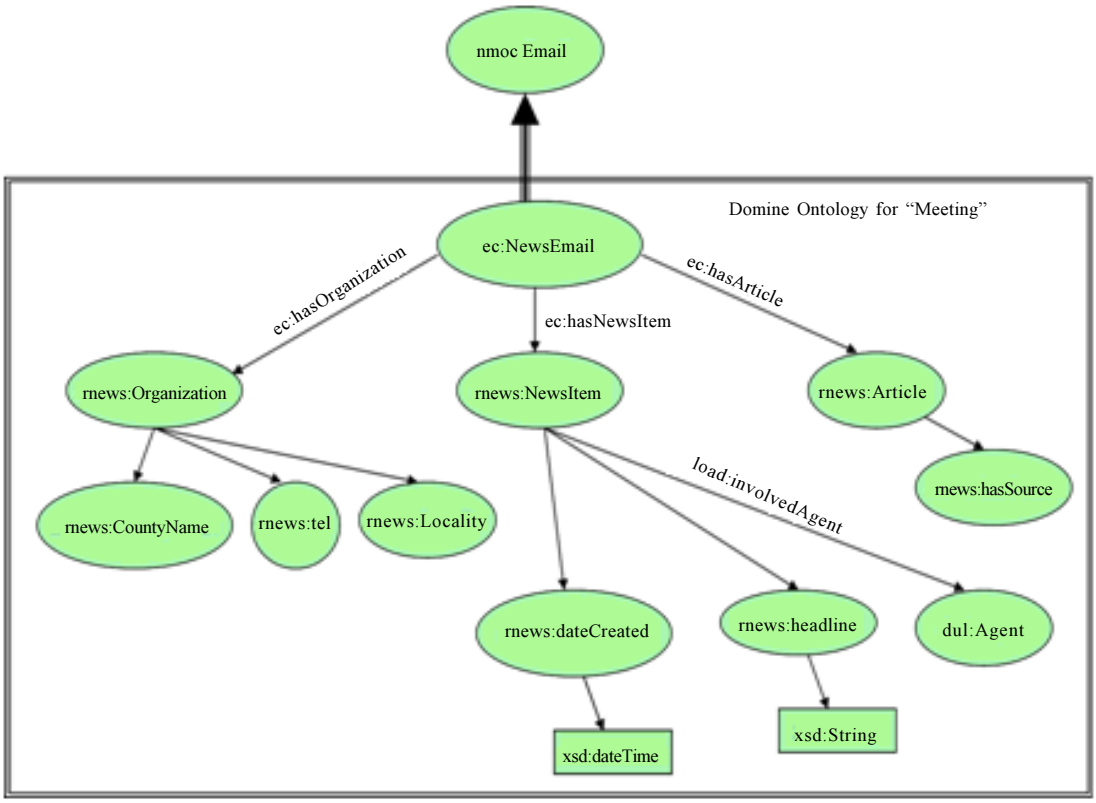

Figure 2. Domain Ontology for Email "*News*"

3. Fuzzy Systems for Reasoning

During the recent past, information systems have experienced signiûcant improvements in intelligent information processing, thereby stimulating advances in the Knowledge Management area. Reasoning is one of the key features of these technologies, with a huge variety of possible application domains and different reasoning techniques for solving each particular problem (Iglesias and Lehmann 2011).

Reasoning over imperfect information, which is inherent to most of the real world application domains, is one of the main research issues in 'Knowledge Management'. In a broad sense, two approaches can be used to deal with non-perfect information. The probabilistic approach is able to deal with the uncertain nature of the information (e.g., modeling sensor accuracy when acquiring data) whereas the fuzzy logic one is able to manage the vagueness of concepts arising from human perception and cognition processes (enabling to formally model, e.g., concepts as '*tall*', '*cheap*', '*easy*', etc.) [12]. This is the reason, fuzzy logic and related techniques can be implemented to resolve the various kinds of vagueness encountered in natural language processing applications. The multilingual emails domain, wherein the language used for communication can be bit informal, has significant ambiguity.

Let's consider an example email from "*Meeting and Schedule*" domain for multilingual emails: "*Najib is meeting Emad at hotel.*

He will then go to watch the football match". In the second sentence, for the purpose of extracting relationships and involved agents, it is not clear who the word "He" is referring to purely from programming aspect; even though as human beings we know that "*He*" refers to Najib. This can be resolved using the context from the previous sentence using the fuzzy logic methodology. In this case, since "*He*" is used as a subject in second sentence, a higher weightage will be given to the Subject, Najib, in the previous sentence.

4. Fuzzy Logic in Natural Language Processing

Natural language is by nature full of "*fuzziness*". We often use fuzzy concepts like "*hot*" or "*tall*" to define objects around us. The traditional applications of science often dealt with strict values which made them less suitable for modelling things which are by nature fuzzy, like language.

Fuzzy logic is thus now being applied to these domains so as to formalize language in order to develop applications, like email ontology system, for it. It can also make systems smarter and mirror the way humans think. For example, in a system if one has to apply the rule: "*If it is getting cooler add some heat*", it is very difficult to mirror this in current systems since one would have to answer questions like how much cooler , from which temperature should we take a reference or how much heat to add etc.

Fuzzy logics deal with incomplete and uncertain information. When we have imprecise, inconsistent and inexact information then the Fuzzy sets are being used.

Fuzzy set A of X is characterized by its membership function:

$A = \mu_A(x)$ and in the unit interval $[0, 1]$

$\mu_A(x) : X \rightarrow [0, 1], x \in X,$ where X is Universe of discourse.

or

$A = \mu_A(x1)/x1 + \mu_A(x2)/x2 + \ldots + \mu_A(xn)/xn$

Where "+" is union

If we take a specific multilingual email ontology domain like: "*Discussion and Comment*" and suppose in a mail the ontology learning system encounters a statement: "*This document was very long and did not bring out the meaning of the topic*". System should resolve '*hasComment*' as long and unclear. In this case the concept of "*long*" is fuzzy, so we need to use fuzzy set in it.

For instance, x is Long; defined as

$Long = \mu_{Long}(x) \rightarrow [0, 1]$, where "*Long*" is fuzzy set.

$Long = 0.56/x1 + 0.6/x2 + 0.65/x3 + 0.67/x4 + 0.69/x5$

In our system, we will apply fuzzy logic to develop a learning model which will help resolve the vagueness of language and help us extract the necessary parameters from an email, say meeting details, by understanding and modelling the way humans write such emails.

5. Fuzzy Logic Email Ontology Learning System

In our system we wish to infer the meaning of the email and extract certain parameters from these emails. We will model this behaviour by incorporating our current algorithm with a traditional fuzzy decision system. The system is as shown in Figure 3 below.

The system requires a large amount of knowledge to be acquired to develop a powerful logic system. We need to develop some procedures to automatically extract information for fuzzy semantic logic [6].

5.1 Email Pre-Processing

In our current algorithm as shown in Figure 1, we are pre-processing our email by the following steps:

• Converting email to XML structure.

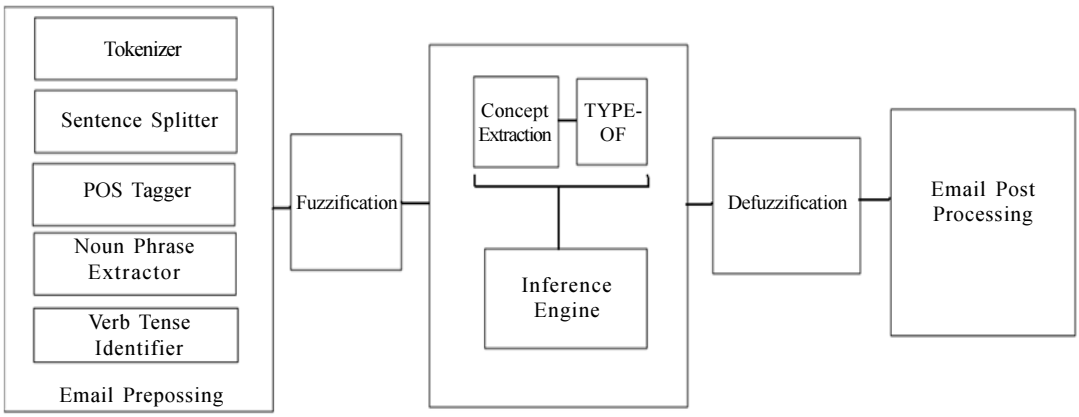

Figure 3. Fuzzy Decision System for Email Ontology Learning

• Dividing email into subsections such as To, CC, Subject, Body etc.

• Subject and Body Processing by

  o Tokenizer

  o Sentence Splitter

  o Part of Speech tagger(POS tagger)

  o Noun Phrase extraction

  o Determining relations

• Determining domain of email through statistical methods.

As described in the Figures for domain ontologies, we need to extract appropriate agents and relations to correctly capture the vocabularies for domain ontologies. This involves a lot of vagueness due to informal nature of email communication as well as vagueness involved in natural languages. For example, consider the statement: "*The professor is planning to come in Seminar Hall. His meeting is scheduled here at* 10:00 *AM*".

In the second sentence {*his}, {here*} are fuzzy, to resolve the fuzziness we used fuzzy logic for the sentence. {*is*} needs to be mapped with the subject of the previous sentence i.e { *professor*}. Fuzzy Logic has been successfully applied to the description of words. Fuzzy logic evaluates whether a recognized word is semantically appropriate with respect to the identified sentence

5.2 Fuzzy Sets

For ontology learning system in multilingual email domain, the most important constructs are noun and noun phrases in the sentences. The vagueness of the pronouns (he, she, they, his, her etc.) needs to be deciphered in order to arrive at correct relations and involved agents for ontology development.

For studying the constructs for the nouns and noun phrases, a detailed statistical study has been performed on three email corpuses:

• Enron Email Corpus

• British Columbia Conversation Corpus (B3C)

• Customized Arabic Email Corpus

From the above three corpus, we selected significant number of emails for our research. We categorized the emails in four different categories. In all, we considered more than 2000 emails for the study. The Chart 1 shows the split of emails collected from different datasets.

Table 1 above shows the split of emails for different domain of emails considered for the analysis:

From the fuzzy sets perspective, two statistical details are obtained from the emails and all the statements within the emails:

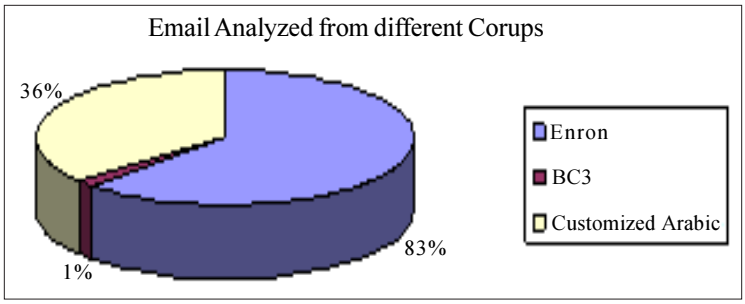

Chart 1. Email Analyzed from different domains

Email Domain	Enron	Bc3	Cusomized Arabic
Meeing and Schedule	345	8	225
News	230	6	125
Discussion and Comments	650	12	350
Collaboative and Technical Support	120	5	80
Total	**1345**	**31**	**780**

Table 1. Classification of Emails from three Email Corpus Dataset

1. Distribution of number of nouns in a sentence.
2. Location of noun within the sentence (Normalized Value)

The probability distribution chart for distribution of number of nouns is shown below:

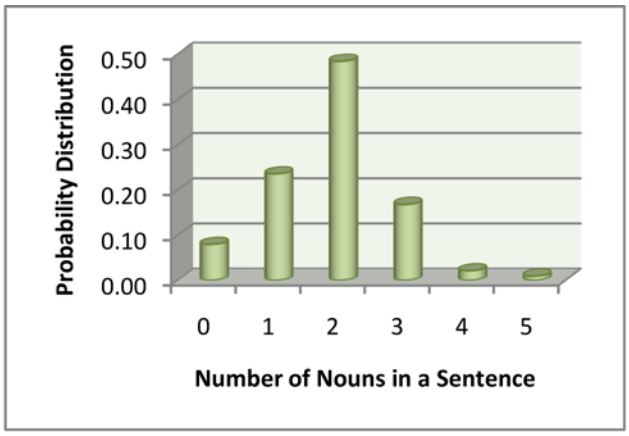

Chart 2. Discrete distributions of nouns in a sentence

The discrete probability distribution is normally distributed. It is then converted into a normal Gaussian distribution function for the purpose of fuzzy sets:

$$f(X) = \sqrt{\frac{1}{2\pi\sigma}} \exp\left[\frac{-(X-\mu)^2}{2\sigma^2}\right]$$

Here σ and μ are respectively standard deviation and mean of our probability distribution function. This distribution function is one of the inputs in the fuzzy application. The corresponding normal guassian distribution for the discrete information on number of nouns in a sentence is shown below:

It is used for the concluding the co-reference for an encountered pronoun. So if in a sentence, we encounter a pronoun, the

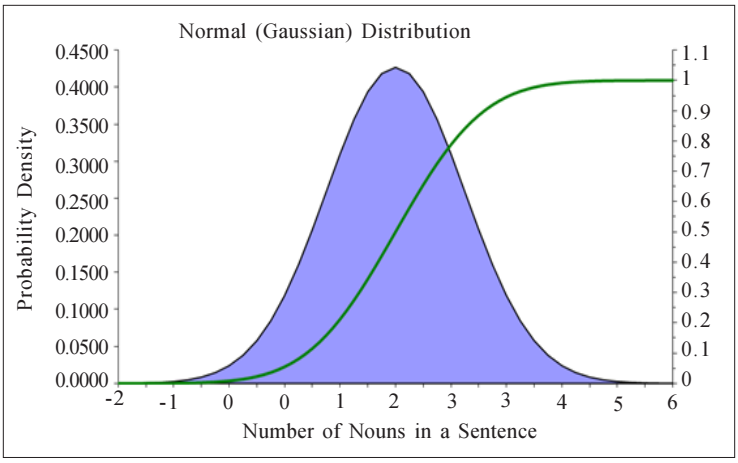

Chart 3. Gaussian Distribution

preceding sentence is used to determine its reference. Since in ontology applications, both the agents and relations among them are important, sentences with two or more nouns are most critical. From the distribution, we assumed a membership function similar to probability distribution. So the chances of matching the pronoun with its actual noun are high when number of nouns are two in the preceding sentence and this decreases as the number of nouns increase or decrease.

The second important aspect in correctly deciphering the reference for the pronoun is location of noun within the preceding sentence. Three locations are considered for the noun in the sentence: Beginning, Middle and End. The triangular membership function is shown below:

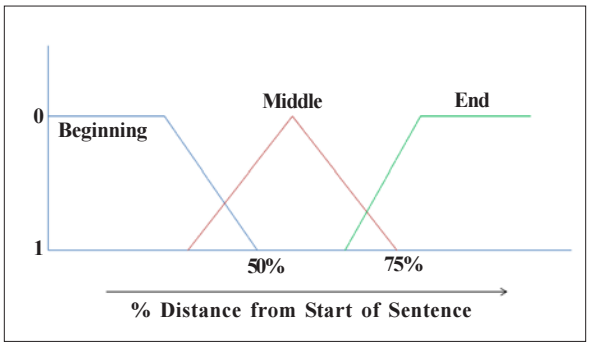

Chart 4. Triangular membership function

The membership function, mathematically are defined as:

$$\text{Beginning} = \begin{cases} 1.0 \; for \; 0\% < distance \leq 33\,\% \\ \dfrac{(50 - distance)}{(50 - 33)} \; for \; 33\,\% < distance \leq 50\,\% \\ 0.0 \; for \; 50\% < distance \end{cases}$$

$$\text{Middle} = \begin{cases} 0.0 \; for \; 0\% < distance \leq 33\,\% \text{ and } 0.0 \; for \; 75\% < distance \\ \dfrac{(distance - 33)}{(50 - 33)} \; for \; 33\,\% < distance \leq 50\,\% \\ \dfrac{(75 - distance)}{(75 - 25)} \; for \; 33\,\% < distance \leq 50\,\% \end{cases}$$

$$\text{End} = \begin{cases} 0.0 \; for \; 0\,\% < distance \le 67\,\% \\ \dfrac{(distance - 67)}{(75 - 67)} \; for \; 67\,\% < distance \le 75\,\% \\ 1.0 \; for \; 75\,\% < distance \end{cases}$$

The decision process in selection of the pronoun co-reference is:

1. GOOD MAPPING: IF number of nouns in the preceding sentence are less than or equal to two AND one of them lie at the beginning of sentence, THEN perform a GNP check (Gender, Number, Person) check and select the first noun for mapping.

2. MODERATELY GOOD MAPPING: IF number of nouns in the preceding sentence are less than or equal to two AND one of them lie at the middle and the other one (if present) at the end, then select the first one.

3. NO MAPPING: All other conditions ignore the pronoun mapping process (like presence of three or more nouns).

These three outputs are crisp values from the fuzzy analysis of the pronoun mapping. Let's consider an example:

The big orange balloon rose high above in the sky. The little boy was looking at it intently.

We need to decipher the reference for "*it*" from the second sentence. There are two nouns in the preceding sentence which are obtained from POS tagger built in GATE: Balloon and Sky. Let's take '*Balloon*' first. Since there are two nouns, the membership function for the number of nouns gives a value of 0.52. The location of 'Balloon' is about 40% in the first statement. So its membership in the set Beginning is 0.59 whereas for the set Middle it is 0.41.

Applying these membership functions on Rule 1 and Rule 2 above, we have:

Good Mapping = Min (0.52, 0.59) = 0.52

Moderately Good Mapping = Min (0.52, 0.41) = 0.41

These values are plotted on the membership function graph of "*Good Mapping*" and "*Moderately Good Mapping*" as shown in graph below:

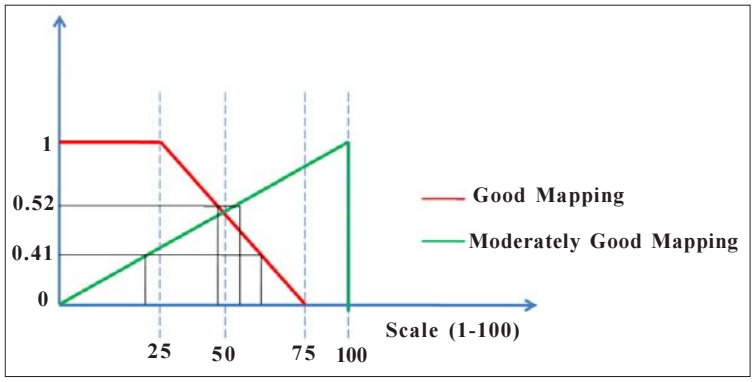

Chart 5. Membership Function for Defuzzification

The defuzzication is then achieved by evaluating the crisp output using the following formula:

$$Output = \frac{\sum_{i=1}^{n} (Centroid \times strength)_i}{\sum_{i=1}^{n} (strength)_i}$$

The crisp output in our case, then is:

$$Crisp\ Output = \frac{29.17 \times 0.52 + 66.67 \times 0.41}{0.52 + 0.41}$$

$$= 45.7$$

At a value of 45.7, Good Mapping prevails, so we go ahead with Gender, Number and Person check and select the first noun for mapping it. Thus we map "*it*" in the second sentence to "*balloon*" in the first sentence with a high confidence that mapping is correct.

6. Conclusion and Future Work

Sometimes language cannot be defined precisely, but depend on the contexts. In that case we can use Fuzzy sets as a knowledge representation. This is especially true in case of email communication in multilingual.

Fuzzy logics are used to make decisions in vague domains. That can be utilized in new ontology technologies for domains with vagueness like multilingual emails. In our system we are trying to resolve the ambiguity of Pronoun. Ambiguities of adjective, verb etc. have been ignored for the present. These will need to be tacked in the future for appropriate development of email ontologies.

The system requires a large amount of knowledge to be acquired to develop a powerful logic system. We need to develop some procedures to automatically extract information for fuzzy semantic logic. We have to collect more information in efficient ways. The ambiguity of discourse based on contextual information is important for understanding of the sentence which decides the precise interpretation of its semantics. A neural network based model is more appropriate, in general, for pattern matching and a fuzzy logic based model is required to remove the ambiguities in the language used in email communication.

References

[1] Clark, J., Gonzales-Bernes, J. (2008). Coreference: Current Trend and Future Directions. *In*: Technical Report, Language and Statistics II Literature Review.
[2] Kosko, Bart. (1997). *Fuzzy Engineering*. Prentice-Hall, Inc.
[3] McNeill, Daniel, Frieberger, Paul. (1993). *Fuzzy Logic*. Simon & Schuster.
[4] Ross, Timothy, J. (2004). *Fuzzy Logic with Engineering Applications.* s.l.: John Wiley & Sons Ltd. 0-470-86075-8.
[5] Lakoff, George. (1973). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*. p. 458-508.
[6] Zadeh, L. A. (1965). Fuzzy Sets, Berkeley : University of California, *Information and Control*, 8, 338 - 353.
[7] Zadeh, Lofti. (1972). A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges. *Journal of Cybernetics*. 2/3, 4-34.
[8] Zadeh, Lofti. (2002). From Computing with Numbers to Computing with Words.*Applied Math and Computer Science*. 12/3
[9] Zhou, L. (2007). Ontology learning: State of the art and open issues. *Information Technology and Management*, (3) 241-252
[10] Zhou, J. Z. L. S. Z., Liang, Y. (2007). Fuzzy Ontology Model for Knowledge Management.
[11] Bontcheva, K., Tablan, V.et al. (2004). Evolving GATE to Meet New Challenges in Language Engineering, *Natural Language Engineering*, 10 (3/4) 349-373.
[12] Iglesias J., Lehmann, J. (2011). Towards Integrating Fuzzy Logic Capabilities into an Ontology-based Inductive Logic Programming Framework.: Intelligent Systems Design and Applications (ISDA), 11th International Conference on.