Fuzzy Set Theory: a Useful Interlingua?

Chris Cornelis[1], Martine De Cock[1], Dick Botteldooren[2], Etienne Kerre[1]

[1] *Fuzziness and Uncertainty Modelling Research Unit*
Department of Mathematics and Computer Science
Ghent University, Krijgslaan 281 (S9), B-9000 Gent, Belgium

[2] *Acoustics Research Group, Department of Information Technology*
Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

{*chris.cornelis,martine.decock,dick.botteldooren,etienne.kerre*}*@rug.ac.be*

Abstract. The practice, common in fuzzy expert systems, to map linguistic data into fuzzy sets has some important reverse implications: given membership functions that represent different terms of a linguistic variable in one natural language X, we can set out to search which of them best matches a given term in another language Y, on the basis of similarity assessments between individual terms. In other words, we are considering fuzzy set theory as an interlingua for use in automated translation. Our approach is applied to data gathered in an International Annoyance Scaling Study.

1 Introduction

The use of linguistic terms, internally represented as membership functions (i.e. fuzzy sets), has always been a predominant feature of fuzzy expert systems. Noticing the need for efficiency and expressivity in membership function construction, researchers have proposed modelling techniques for this purpose (see e.g. [3]). Several approaches rely on *psycholinguistic experiments*, by which the construction process can be made to yield membership functions that reflect, in a mathematical sense, what a given concept "means" to a representative group of individuals.

Given membership functions that represent different terms of a linguistic variable, we may study how similar they are. The notion of similarity is well–known in fuzzy set theory, and a panoply of measures exist that assess numerically how accurately two fuzzy sets resemble each other. Going one step further, we can set out to search, by the same mechanism, for the linguistic term in language X that best matches a given term in another language Y. In this way, fuzzy set theory may effectively aid, as an interlingua independent of natural language, to resolve some of the ambiguities that arise in the domain of automated translation.

Our work is based, on one hand, on a corpus of data gathered in an International Annoyance Scaling Study [2], conducted in 9 modern languages, where the relationship between 21 different linguistic terms and their corresponding noise annoyance level was under survey, and on membership construction techniques [3] on the other hand. Specifically, we apply similarity measures to match modified terms with accurate translations.

2 Membership Functions

In fuzzy set theory [5] it is acknowledged that the transition between belonging and not belonging to a set can be gradual instead of abrupt; in other words that an object can belong to a set to a certain degree. Therefore a fuzzy set A on a universe X is characterized by a $X \rightarrow [0,1]$ mapping, for simplicity also denoted by A, which is called the membership function of A. Hence for all x in X, $A(x)$ is the degree to which x belongs to the fuzzy set A. By $\mathcal{F}(X)$ we denote the class of fuzzy sets on X. It is well–known that this graded approach makes fuzzy set theory a very suitable framework for the mathematical representation of vague linguistic terms. In this section we briefly discuss the construction of membership functions for noise annoyance terms.

As an environmental factor, noise has several adverse effects on men. Annoyance or disturbance is commonly used as an impact indicator of these effects. To get a better understanding of what people really mean when they use some annoyance terms such as “not at all annoyed”, “slightly annoyed”, “very annoyed”,... an International Annoyance Scaling Study was conducted [2]. People were asked to indicate with a mark on a line what each term meant to them, with the most left-hand side being no annoyance at all and the most right-hand side being the most possible degree of annoyance one can imagine. While processing the data, each mark on the line was converted into the distance (expressed in centimetres) from the left-hand side on the 10 centimetres long line. This resulted in a continuous numerical domain ranging from 0 to 10 and a dataset containing 21 values for each subject. It should be stressed that these data were not gathered with the purpose of membership construction in mind.

When an informant places a mark for a particular term A we assume that he means: “this (and the surrounding) annoyance level(s) I call A, but the other ones not”. Following the approach of [3], we compute the average number of marks taken over all informants, which corresponds to constructing the (normalized) probabilistic histogram, and treat this as a membership function for A. Optionally, we could fit an exponential bell–shaped curve on the histogram, but we feel that in this case the fitting process might hide some of the small characteristics that could be important to distinguish between terms with a close meaning. Therefore we will focuss on the unfitted membership functions. For efficiency in comparison between fuzzy sets we have rounded the numerical data before starting the construction of the membership functions: hence we only need to compare membership values in 10 points for every fuzzy set. Figure 1 depicts the concepts “presque pas”, in French, and “very”, in English.

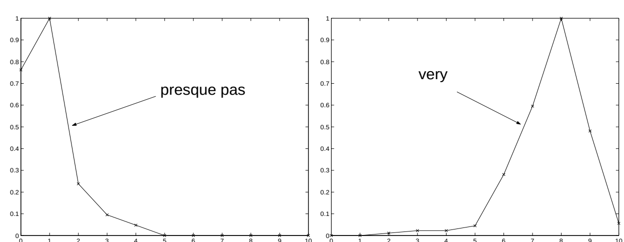

Figure 1: Membership functions for French “presque pas” and English “very”

3 Comparing Membership Functions

Generally speaking, a similarity measure on a universe X is a [0,1]–valued indicator suitable for the comparison of fuzzy sets on X, i.e. a binary fuzzy relation on $\mathcal{F}(X)$. Depending on the requirements imposed on the measures, different indicators with varying behaviour can be obtained. Following Tsiporkova and Zimmermann [4], we make a basic distinction between measures inspired by set equality, and degrees of compatibility or overlap: a binary fuzzy relation Eql on $\mathcal{F}(X)$ is called a T–equality if $Eql(A,B) = 1 \iff A = B$, $Eql(A,B) = Eql(B,A)$ and $T(Eql(A,B), Eql(B,C)) \leq Eql(A,C)$ are satisfied for any fuzzy sets A, B and C on X, where T is any t–norm, i.e. a symmetric, associative, increasing $[0,1] \times [0,1] \to [0,1]$ mapping satisfying $T(1,x) = x$ for every $x \in [0,1]$. A reflexive, symmetric binary fuzzy relation Com on $\mathcal{F}(X)$ is called a degree of compatibility if it satisfies the condition $Com(A,B) = 0 \iff \sup\limits_{x \in X} \min(A(x), B(x)) = 0$, for any $A, B \in \mathcal{F}(X)$.

Whether one measure is preferable over another depends strongly on the application at hand, and most often it is worthwile to take into account the results from different similarity assessments and optionally aggregate them into a single value. A directly applicable class of T–equalities is given by, for A and B in $\mathcal{F}(X)$:

$$Eql_T(A,B) = T\left(\inf_{x \in X} \mathcal{I}_T(A(x), B(x)), \inf_{x \in X} \mathcal{I}_T(B(x), A(x))\right)$$

where the residual implicator $\mathcal{I}_T$ is defined by, for x and y in $[0,1]$ as $\mathcal{I}_T(x,y) = \sup\{\gamma \in [0,1] | T(x,\gamma) \leq y\}$. These mappings are studied in detail in [1]. Intuitively, it makes sense to apply them, since concepts that are linguistically similar will result in approximately equal membership functions.

As adequate examples of degrees of compatibility, we quote the following measures[1] $\mathcal{S}_1$ and $\mathcal{S}_2$ for a finite universe X:

$$\mathcal{S}_1(A,B) = \frac{\sup\limits_{x \in X} T(A(x),B(x))}{\sup\limits_{x \in X} S(A(x),B(x))} \quad \mathcal{S}_2(A,B) = \frac{\sum\limits_{x \in X} T(A(x),B(x))}{\sum\limits_{x \in X} S(A(x),B(x))}$$

Guided by the observation of Zwick et al. [6], who empirically compared the performance of overlap degrees, that “good” measures concentrate their attention on a single value rather than performing some sort of averaging, we might prefer $\mathcal{S}_1$ over $\mathcal{S}_2$. Indeed, $\mathcal{S}_1$ compares the peak regions of both fuzzy sets by assessing the height of their intersection. While $\mathcal{S}_1$ will in general assign low degrees of similarity to sufficiently different concepts, it might however not distinguish very well between concepts with close, but not exactly the same meaning. An example at hand is the following problem: suppose we have to translate the German term “mittelmäßig” (meaning, literally, “around the middle”) into English. From picture 2a), it is clear that $\mathcal{S}_1$–comparison with both “fairly” and “moderately” will result in similarity degree 1, which is counter–intuitive.

Measure $\mathcal{S}_2$ will notice the difference, preferring “moderately” (degree 0.73) to “fairly” (degree 0.34), and warning the user, through the occurence of a highest similarity degree still considerably less than one, to proceed with caution because this concept is apparently not straightforward to translate exactly; we could not find a fuzzy set for an English concept with higher degree of similarity to “mittelmäßig”. On the other hand, measure $\mathcal{S}_2$ might

[1] T denotes a t–norm and S is a t–conorm, i.e. a symmetric, associative, increasing $[0,1] \times [0,1] \to [0,1]$ mapping satisfying $S(0,x) = x$ for every $x \in [0,1]$.

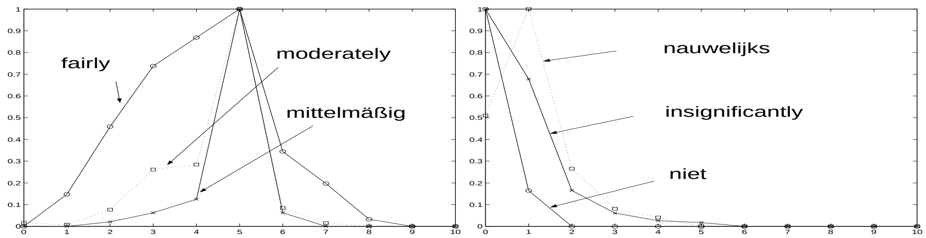

Figure 2: a) Comparing “mittelmäßig” with fairly and moderately b) Comparing “insignificantly” with “niet” and “nauwelijks”

optimistically assign a high degree of similarity to fuzzy sets with considerable overlap but having their peaks quite apart. Figure 2b) provides an example of this phenomenon: “niet” and “nauwelijks” in Dutch evaluate to the same similarity degree 0.6 when compared with the English “insignificantly”, but a peak comparison makes it clear that “niet” is in fact a better translation.

Basing ourselves on the findings of [4], where ways of combining the best of different worlds into a robust similarity indicator are outlined, we opted for the following generic hybrid measure (it is noted that several variations exist on this theme):

$$Sim(A, B) = \min(Com_1(A, B), \max(Eql(A, B), Com_2(A, B)))$$

where Com_1 and Com_2 are degrees of overlap while Eql is a T–equality for a t–norm T.

We will illustrate the general translation process by listing some test results comparing the English modifier “moderately” to some available Dutch alternatives (only the most relevant terms, namely those that are linguistically close, are considered). $\mathcal{S}_1$ and $\mathcal{S}_2$ will play the role of Com_1 and Com_2, while we put $Eql = Eql_W$, where the Łukasiewicz t–norm W is defined as, for x and y in $[0, 1]$: $W(x, y) = \max(0, x + y - 1)$.

	$\mathcal{S}_1$	$\mathcal{S}_2$	Eql_W	$\min(\mathcal{S}_1, \max(\mathcal{S}_2, Eql_W))$
enigszins	0.28	0.25	0.00	0.25
matig	1	0.44	0.31	0.44
tamelijk	0.76	0.02	0.26	0.26
behoorlijk	0.78	0.28	0.00	0.28

The obtained values reflect quite well what we can see in figure 3: of the available Dutch terms, “matig” matches “moderately” best and is thus the obvious candidate for translating “moderately”. But the picture also reveals that, judging by its really wide membership function, “matig” is interpreted in substantially differing ways by the Dutch informants (indeed, a non–negligible part of them regard it as a weakening rather than as an averaging concept), whereas for the English concept the ambiguities are less pronounced. It might also be argued that the impact of Eql_W is usually cancelled by the higher values that $\mathcal{S}_1$ and $\mathcal{S}_2$ yield. This is due to the non–compensatory nature of $\min$ and $\max$, and can be mended by generalizing them to t–norms and t–conorms, respectively, as is a common practice in fuzzy set theory.

4 Conclusion and Final Remarks

In this paper we have demonstrated that the modelling of linguistic terms by fuzzy sets has even more to offer than their use as a representation of knowledge in fuzzy systems. Indeed,

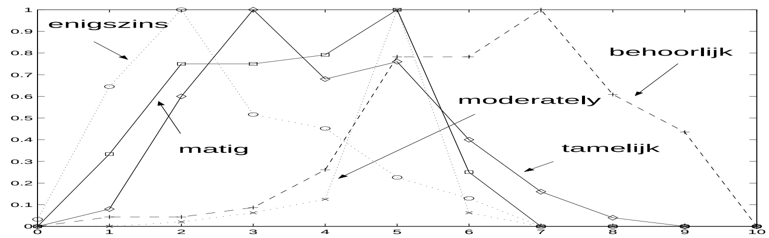

Figure 3: Translating the concept "moderately" into Dutch

representing terms in different languages with the same kind of fuzzy sets makes fuzzy set theory act as a universal interlingua. The quality of the translation depends heavily on the choice of good similarity measures, but they are available.

Constructing suitable membership functions for all linguistic terms in all languages would be an enormous job. Nevertheless fuzzy sets can assist native speakers of different languages explaining to each other the "subtleties" of difficult terms. We see at least two important applications for this procedure:

- Given a fuzzy system with a linguistic interface towards the user, our technique might, in translating the linguistic labels for the system variables, reveal anomalies that a simple dictionary check might overlook.
- In designing adequate terminology for international (technical) projects and standards, it can help pinpoint good common terms (those which are easy to translate).

Acknowledgements

Chris Cornelis and Martine De Cock would like to thank the Fund for Scientific Research Flanders (FWO) for funding the research elaborated on in this paper.

References

[1] B. De Baets, R. Mesiar, Metrics and T–equalities, to appear in: Journal of Math. Anal. Appl

[2] U. Felscher–Suhr, R. Guski, R. Schuemer, Some results of an international scaling study and their implications for noise research, Proceedings of Noise Effects (1998), 733–736.

[3] H. Hersh, M. Caramazza, A fuzzy set approach to modifiers and vagueness in natural language, Journal of experimental psychology **105(3)** (1976), 254–276.

[4] E. Tsiporkova, H.-J. Zimmermann, Aggregation of compatibility and equality: a new class of similarity measures for fuzzy sets, Proceedings of IPMU-Information Processing and Management of Uncertainty in Knowledge-based Systems Seventh International Conference, Paris (1998), 1769–1776.

[5] L. A. Zadeh, Fuzzy Sets, Information and Control **8** (1965), 338–353.

[6] R. Zwick, E. Carlstein, V. Budescu, Measures of similarity among fuzzy concepts: a comparative analysis, International Journal of Approximate Reasoning **1** (1987), 221–242.