Abstract on *Fuzzy information theory with applications to computational linguistics*

Author: Laura Franzoi
Supervisor: Prof. Liviu P. Dinu

Bucharest, 2018

Contents

Introduction and motivation

According to Ethnologue, cf. [7], there are around 7000 living natural languages in the world, and one of the most interesting topic (not only in the academic field, but also in the general public) is their classification. If till 90s the main methods used for classifying the natural languages were the comparative one, the last decades bring us an increasing number of computational approaches for estimating evolutionary histories of languages. The large majority of methods employed in computational historical linguistics used the lexical items, while syntactic approaches were much less present in language classification. Moreover in language classification fuzzy tools and information theory were used only quite sparsely, cf. [5, 11].

In the thesis we have presented new methods for language classification by investigating original mathematical and computational tools to deal with vague (fuzzy) data, irrelevant features (i.e. features which should not contribute to the classification) and even inconsistent features (which do not make sense for specific languages). To evaluate our methods, we have tested them on two significant data sets, one of Ž. Muljačić, cf. [18], and the other one more recent due to G. Longobardi, cf. [3].

It turns out that, to achieve our goals, we had to use tools and ideas derived from Shannon's information theory: consequently, we had to establish and clarify the deep relationship between traditionally distinct approaches to the management of incomplete knowledge, probability/statistics versus possibility/fuzziness.

One of the main ingredients of this thesis is *fuzzy logic*, a hot topic nowadays. The concept of fuzzy set, despite its many applications, has become a battlefield of conflicting opinions, since its first appearance in the article "*Fuzzy Sets*" written by late Lotfi Asker Zadeh [1927-2017] in the journal "*Information and Control*" in 1965, cf. [30]. When he introduced fuzzy sets, he was motivated by the lack of any existing mathematical framework that could meet the complexity of what he called *animate systems*. The notion of a fuzzy algorithm for parking a car was a convincing illustration of this

view, cf. [16]. Today we would first think of *vagueness* and its management, as opposed to *uncertainty* dealt with in probability theory.
Zadeh did not suggest a new theory that wipes out the already known mathematical concepts, but rather an extension of them.
Fuzzy sets have been encountering an increasing interest, and are gaining more and more recognition. Journals and books on fuzzy set theory are published, conferences on fuzzy modeling and control are organized and software packages for fuzzy control are released.

Other powerful tools used in this thesis come from the Information Theory field, which was "officially" born in 1948, cf. [27]; actually Cl. Shannon had already written, during the second world-war, a path-breaking paper on cryptography, inclusive of information-theoretic cryptography, which was declassified, revised and published publicly only in the following year, cf. [28]. The information-theoretic notion of string *distinguishability*, or complementarily of string *confusability*, which is quite central to our work below, pertains originally to "zero-error information theory", cf. [29], and was later generalized to quite a broad framework in [2, 1, 4, 14, 15, 21, 19, 20, 22]. In this work we intend to show that this Shannon-theoretic notion, not to be confused with distances and dissimilarities, is of basic importance in natural language processing and mathematical linguistics. Actually, linguistic evolution sorely needs this notion, as we shall argue below. The intersections of linguistic with the information-theoretic notions of channel noise and channel decoding are quite relevant in this work, and in our opinion leave ample room for future research.

Thesis structure

A general idea of the approach taken in this thesis is given below, relying basically on the introductions to the single chapters.

Chapter 1 discusses the alternative definitions of fuzzy numbers and fuzzy quantities one can find in the literature (inclusive of literature to which this writer has contributed too, cf. [9, 10, 23]) and provides a brief overview of the fundamentals of possibility theory and fuzzy logic which are necessary for the formulation of the problems described in the following chapters. We stress the reasons why we think that having a solid basis for fuzzy arithmetic is essential to properly use fuzzy tools, as needed here in a linguistic and information-theoretic context.

Chapter 2 deals with the notion of irrelevance, which was first put forward jointly by this writer and which is a basic and convenient tool to speed up computations in the arithmetic of interactive fuzzy numbers. After introducing our discussion of irrelevance by tackling fuzzy n-poles, i.e. fuzzy quantities with a finite support of size n, we shall move to more usual continuous supports as are intervals. At the end of the chapter we shall try to understand what happens if the fuzzy quantities one is considering are incomplete, or sub-normal, that is if one allows that a fuzzy quantity is "cut" at a height h which is less than 1. The results are based on papers co-authored by the writer, cf. [9, 24, 10, 12, 25, 11].

Chapter 3 can be seen as a vindication of our claim that fuzzy arithmetic is indispensable for a correct comprehension and use of fuzzy notions and fuzzy tools. In standard information theory input signals are assumed to be crisp, but this is not always realistic. Our approach in this chapter is mixed, fuzzy/stochastic. We argue that fuzziness is an adequate choice when one cannot crisply control each signal fed to the channel. Using the fuzzy arithmetic of interactive fuzzy numbers, we explore the impact that a fuzziness constraint has on channel capacity. The material presented in this chapter relies heavily on our joint published paper [13].

Chapter 4 is an introductory chapter, which paves the way to the dichotomy distance/distinguishability, and so to our applications of fuzziness to computational linguistics and linguistic classification. As we shall comment upon, especially in Chapter 5, this Shannon-theoretic dichotomy is quite basic not only in coding theory, but also in computational linguistics and natural language processing.

Chapter 5 is the central chapter of the thesis from a strictly information-theoretic point of view. It is based mainly on the joint paper [1]. We shall show that the fuzzy and possibilistic approach and the probabilistic approach to information and coding theory are closely related and can be linked by the unifying notion of upper probability.

In **Chapter 6** we provide language classification based on Ž. Muljačić data. In his classifications Ž. Muljačić used vague (fuzzy) data. Using his idea we will compute the fuzzy Hamming distance (that we call Muljačić distance). Using then the Cl. Shannon notion of codeword distinguishability, we will introduce the fuzzy Hamming distinguishability derived from Muljačić distances, which is very easy to compute, and we will call it Muljačić distinguishability. In this chapter we will retake the language classification based on Muljačić distances and extended to Muljačić distinguishability. In the last section of the chapter we will enrich the logical framework and replace minima and maxima with more general t-norm (conjunction) and the corresponding t-conorm (disjunction) as used in the fuzzy community. The distances we will obtain end up supporting the original Muljačić classification.

In **Chapter 7** we re-take Muljačić data with ill-defined features, but as recently suggested by the linguist G. Longobardi, we take into account the fact that some features can be irrelevant (they do not have any sense). For this classification we resort to the Steinhaus transform, a powerful tool which, as we show, allows one to deal with linguistic data which are not only fuzzy, but possibly also irrelevant or even logically inconsistent. The material presented in this chapter is only partly published, cf. [6, 15]. In the last section of the chapter we exhibit a metric distance which allows one to deal jointly with both irrelevance and inconsistency, and which is based on the generalized Steinhaus transform. In a final remark we comment why our metric distances, even if they do not outperform Longobardi's non-metric distances as far as linguistic classification is concerned, might prove useful in situations where his non-metric distances *cannot* be used, and so might contribute to his ambition projects on linguistic evolution.

In **Chapter 8** we review and comment the results obtained in the thesis.

Contributions

This doctoral thesis is based on results [1, 6, 11, 8, 14, 13, 15, 25, 26, 24, 23] published by this writer (some in co-authorship), which are below extended, re-organized and improved. The full list of publication is given also separately at page 9 and 10.

In the thesis we have first discussed the alternative definitions of fuzzy numbers and fuzzy quantities one can find in the literature. The definition we have proposed in the first chapter appears in [25], and we have presented it at the 16th International Conference IPMU in Eindhoven in 2016.

Then we have discussed the different tools one can use to speed up computations using fuzzy quantities. In particular we have considered the discrete case: we have presented the results at the 14th International Conference IPMU in Catania in 2012. After that we have considered the continuous case dealing first with *partial irrelevance*, i.e. the results of an algebraic operation are the same for some distributions, cf. [12], and then with *total irrelevance*; these results are published in [12] and [25] respectively. In particular the writer presented the last at the 16th International Conference IPMU in Eindhoven in 2016.
Then we have noticed that sometimes one can deal with irrelevance in *incomplete* fuzzy arithmetic, i.e. we do not deal necessary with a fuzzy number, but with a *fuzzy quantity*; we can face this situation using the Mamdani implication, or the product between fuzzy numbers with a not pure support. A question arises spontaneously: dealing with generalized and incomplete fuzzy quantities, instead of well-behaved fuzzy numbers, can we yet speak of irrelevance? The writer has published [11] and presented it at the 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computin SYNASC in Timişoara in 2016.

To stress that fuzzy arithmetic is indispensable for a correct comprehension and use of fuzzy notions and fuzzy tools we have introduced the notion of *fuzziness error*, as a sort of counterpart to the standard probabilistic notion of *decoding error*; the results have been published, cf. [13], and presented at the conference Soft Methods for Data Science SMPS in Rome in 2016.

We have proved that the fuzzy and possibilistic approach and the probabilistic approach to information and coding theory are closely related and can be linked by the unifying notion of *upper probability.* To do that we have considered and proved two inverse problems on coding theory. The results have been published in Fundamenta Informaticae in 2015, cf. [1].

After this introduction in Chapters 6 and 7 we have presented linguistic classification results using fuzzy and information theory tools.
To start the classification we have noticed that back in 1967 the Croat linguist Ž. Muljačić had used a fuzzy generalization of the Hamming distance between binary strings to classify Romance languages, cf. [18]. Just two years before L. A. Zadeh had introduced the notion of fuzziness. In 1956 Cl. Shannon had introduced the notion of codeword distinguishability in zero-error information theory. Analyzing all these facts we have noticed that distance and distinguishability are subtly different notions, even if, with distances as those usually met in coding theory, the need for string distinguishabilities evaporates, since the distinguishability turns out to be an obvious and trivial function of the distance. We have derived fuzzy Hamming distinguishabilities from Muljačić distances. The results were published in [14] and presented at FUZZ-IEEE in Naples in 2017. Then we asked ourselves: what happens if we change the logical framework and instead of considering the classical conjunction and disjunction, one considers T-norms and T-conorms? The found results have been published in [15], and presented at the 21st International Conference KES-2017 in Marseille in 2017. With this publication we have presented a first classification of Romance languages using Muljačić data.

It could happen, otherwise, that in language classification some features are absent in two or more languages (they are irrelevant). Do these features contribute in the language classification? Using a powerful mathematical tool called Steinhaus transform, we have tried to understand what happen in the language classification and we have found that the results perform quite well the already existing classifications we can find in the literature. The results presented in the thesis have been published, cf. [8], and presented at the 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing SYNASC in Timişoara in 2017.

Finally we have focus on data due to G. Longobardi's school, which are involved in an ambitious and innovative project on language classification based on syntax, cf. [3, 17]: languages are represented through yes-no strings of length 53, each string position corresponding to a syntactic feature which can be present or absent. However, due to a complex network of logical implications which constrain features, some positions might be undefined (logically inconsistent). Using also in this case the Steinhaus transform of Hamming-like distances we have presented a new classification. While G. Longobardi uses a distance which is not metric, we have found his same classification using a metric. The results have been published, cf. [6], and presented at 17th International Conference IPMU in Cádiz in 2018.

Finally we have asked ourselves if we can find a mathematical tool that allows us to consider irrelevance and inconsistency not separately. The answer is the generalized biotope transform. These last results are, however, not published yet but we intend to submit them soon for publication and we are confident that they will prove to be quite a valuable tool in linguistic classification and in linguistic phylogeny.

List of publications

This doctoral thesis is based on the following published results by this writer:

1. A. Sgarro, L. Franzoi, *Livre flou*: `www.dmi.units.it/~sgarro/livre_flou.pdf`

2. A. Sgarro and L. Franzoi, *Fuzzy Arithmetics for Fuzzy n-Poles: When is Interactivity Irrelevant?*, IPMU (3) 2012, pp. 1 – 8, doi: 10.1007/978-3-642-31718-7, ISSN: 1865-0929, Springer Berlin Heidelberg

3. L. Franzoi, A. Sgarro, *(Ir)relevance of interactivity in fuzzy arithmetic, Mathematica Pannonica*, 25/1 (2014), pp. 1 – 11

4. L. Bortolussi, L. P. Dinu, L.Franzoi, A. Sgarro, *Coding theory: a general framework and two inverse problems*, (2015), Fundamenta Informaticae, vol. 141, p. 297 – 310, ISSN: 0169-2968, doi: 10.3233/FI-2015-1277

5. A. Sgarro and L. Franzoi, *(Ir)relevant T-norm joint distributions in the arithmetic of fuzzy quantities*, (2016), IPMU, International Processing and Management of Uncertainity in Knowledge – Based Systems, vol. 611, pp. 3 – 11, doi: 10.1007/978-3-319-40581-0_1, ISSN 1865-0929

6. L. Franzoi and A. Sgarro, *Fuzzy signals fed to Gaussian channels*, (2016), Soft Methods for Data Science, vol. 456, pp. 221 – 228, doi: 10.1007/978-3-319-42972-4_28

7. L. Franzoi, *Irrelevance in incomplete fuzzy arithmetic*, (2016) SYNASC (Symbolic and Numeric Algorithms for Scientific Computing), pp. 287 – 291, ISSN: 2470-881X, doi: 10.1109/SYNASC.2016.052

8. L. Franzoi, A. Sgarro, *Fuzzy Hamming distinguishability*, in press in 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), doi: 10.1109/FUZZ-IEEE.2017.8015434

9. L. Franzoi, A. Sgarro, *Linguistic classification: T-norms, fuzzy distances and fuzzy distinguishabilities*, (2017) – Procedia Computer Science 112 (2017) 1168 – 1177 – 21th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES 2017, 6-8 September 2017, Marseille, France,
doi: 10.1016/j.procs.2017.08.163

10. L. Franzoi, *Jaccard – like variation of the fuzzy Hamming distance*, (2017) – SYNASC 2017 – 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017, Timisoara, Romania, September 21-24, 2017. pp. 196 - 202

11. A. Dinu, L. P. Dinu, L. Franzoi, A. Sgarro, *Steinhaus transforms of fuzzy string distances in computational linguistics*, (2018) – IPMU, International Processing and Management of Uncertainity in Knowledge – Based Systems, vol. 853, pp. 171 – 182, doi: 10.1007/978-3-319-91473-2, ISBN 978-3-319-91472-5

Bibliography

[1] L. Bortolussi, L. P. Dinu, L. Franzoi, and A. Sgarro. “Coding Theory: A General Framework and Two Inverse Problems”. In: *Fundam. Inform.* 141.4 (2015), pp. 297–310.

[2] L. Bortolussi, L. P. Dinu, and A. Sgarro. “Spearman permutation distances and Shannon’s distinguishability”. In: *Fundamenta Informaticae* 118.3 (2012), pp. 245–252.

[3] L. Bortolussi, A. Sgarro, G. Longobardi, and C. Guardiano. “How many possible languages are there?” In: *Biology, Computation and Linguistics* (2011), pp. 168–179.

[4] N. Chomsky. “On the Nature of Language”. In: *Origins and Evolution of Language and Speech* Annals of New York Academy of Sciences (1976), pp. 46–57.

[5] A. M. Ciobanu, L. P. Dinu, and Sgarro A. “Towards a Map of the Syntactic Similarity of Languages”. In: *CICLing 2017* LNCS 10761 (2018), pp. 1–15.

[6] A. Dinu, L. P. Dinu, L. Franzoi, and A. Sgarro. “Steinhaus transforms of fuzzy string distances in computational linguistics”. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part I.* 2018, pp. 171–182.

[7] *Ethnologue.*
`https://www.ethnologue.com/`. 2018.

[8] L. Franzoi. “Jaccard-like fuzzy distances for computational linguistics”. In: *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017, Timişoara, Romania, September 21-24, 2017*, pp. 196–202.

[9] L. Franzoi. *Quadrati di numeri fuzzy.* Unpublished manuscript, Università degli Studi di Trieste, Italia, 2010.

[10] L. Franzoi. *Interactivity in fuzzy arithmetic.* Unpublished master’s thesis, Università degli Studi di Trieste, Italia, 2012.

[11] L. Franzoi. "Irrelevance in Incomplete Fuzzy Arithmetic". In: *18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2016, Timişoara, Romania, September 24-27, 2016*. 2016, pp. 287–291.

[12] L. Franzoi and A. Sgarro. "(Ir)relevance of interactivity in fuzzy arithmetic". In: *Mathematica Pannonica* 25/1 (2015), pp. 1–11.

[13] L. Franzoi and A. Sgarro. "Fuzzy Signals fed to Gaussian Channels". In: *Soft Methods for Data Science, SMPS 2016, Rome, Italy, 12-14 September, 2016*. 2016, pp. 221–228.

[14] L. Franzoi and A. Sgarro. "Fuzzy Hamming distinguishability". In: *2017 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2017, Naples, Italy, July 9-12, 2017*. 2017, pp. 1–6.

[15] L. Franzoi and A. Sgarro. "Linguistic classification: T-norms, fuzzy distances and fuzzy distinguishabilities". In: *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference KES-2017, Marseille, France, 6-8 September 2017*. 2017, pp. 1168–1177.

[16] A. Goldstein. *Oral-History: Lotfi Zadeh.* `http://www.ieeeghn.org/wiki/index.php/Oral-History:Lotfi_Zadeh`.

[17] G. Longobardi, A. Ceolin, L. Bortolussi, C. Guardiano, M. A. Irimia, D. Michelioudakis, N. Radkevich, and A. Sgarro. "Mathematical modeling of grammatical diversity supports the historical reality of formal syntax". In: *University of Tübingen, online publication system Tübingen DEU* (2016), pp. 1–4.

[18] Ž. Muljačić. "Die Klassifikation der romanischen Sprachen". In: *Rom. Jahrbuch 18* (1967), pp. 23–37.

[19] V. Novák. "Fuzzy Logic in Natural Language Processing". In: *2017 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2017, Naples, Italy, July 9-12, 2017*. 2017.

[20] A. Sgarro. "Possibilistic information theory: a coding-theoretic approach". In: *Fuzzy Sets and Systems* 132(2) (2002), pp. 11–32.

[21] A. Sgarro and L. Bortolussi. "Codeword distinguishability in minimum diversity decoding". In: *Journal of Discrete Mathematical Sciences and Cryptography* 9:3 (2006), pp. 487–502.

[22] A. Sgarro and L. P. Dinu. "Possibilistic entropies and the compression of possibilistic data". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002), pp. 635–653.

[23] A. Sgarro and L. Franzoi. *Livre Flou.* `http://www.dmi.units.it/~sgarro/livre_flou.pdf`.

[24] A. Sgarro and L. Franzoi. "Fuzzy Arithmetics for Fuzzy n-Poles: When Is Interactivity Irrelevant?" In: *Advances in Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy, July 9-13, 2012, Proceedings, Part III.* 2012, pp. 1–8.

[25] A. Sgarro and L. Franzoi. "(Ir)relevant T-norm Joint Distributions in the Arithmetic of Fuzzy Quantities". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 16th International Conference, IPMU 2016, Eindhoven, The Netherlands, June 20-24, 2016, Proceedings, Part II.* 2016, pp. 3–11.

[26] A. Sgarro and L. Franzoi. "(Ir)relevant T-norm joint distributions in the arithmetic of fuzzy quantities". In: *IPMU, International Processing and Management of Uncertaintiy in Knowledge - Based Systems.* Vol. 611. 2016, pp. 3–11. DOI: 10.1007/978-3-319-40581-0_1.

[27] C. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27 (1948), pp. 379–423,623–656.

[28] C. Shannon. "Communication Theory of Secret Systems". In: *The Bell System Technical Journal* 28 (1949), pp. 656–715.

[29] C.E. Shannon. "The zero-error capacity of a noisy channel". In: *IRE Transactions on Information Theory* 2 (1956), pp. 8–19.

[30] L. A. Zadeh. "Fuzzy Sets". In: *Information and Control* 8.3 (1965), pp. 338–353.