Controlling an SMS Transcription System using Heuristic and Empirical Criteria

Grégory Smits[1] and Christine Chardenon[2]

[1]GREYC-University of Caen, Caen France
greg.smits@gmail.com
[2]Orange Labs, Lannion France
christine.chardenon@orange-ftgroup.com

Abstract. Analysis modules which compose a linguistic process often have to cope with the problem of concurrent results generation. Control strategies aim at identifying the most relevant results among all generated ones. Using a generic control approach based on a multicriteria decision aid method, this article presents how empirical and heuristic criteria are combined to improve a SMS transcription system.

Keywords. Natural language processing, multicriteria decision aid, control strategy, sms transcription

1 Introduction

The use of communication devices like mobile phones or computers has lead to the emergence of new means of written spontaneous communication. Simple Message Service (SMS) or even forum on Internet have contributed to the development of a "SMS language". The transcription of text written in "SMS language" into a "standard" language like French is an important issue especially for application like text vocalisation or indexing.

TiLT [1] is a "generic" Natural Language Processing (NLP) toolbox that has been developed to answer different applicative needs, and which has already been applied to various tasks like query correction and indexation, coreference resolution, translation, abridging, ... This NLP system is based on the sequential application of analysis modules which are associated to linguistic resources.

Recently, this toolbox has been adapted to perform SMS transcription for French. This particular use of TiLT has brought concern on a recurrent problem that affects most of the NLP system: concurrent and erroneous results generation and propagation. Indeed, due to imprecisions in the linguistic resources, the inherent ambiguity of natural languages and the lack of complementarity of modular and sequential processes, indeterminations appear at different steps of the analysis process. These indeterminations are characterized by the generation of concurrent results. Some of these indeterminations are legitimate, when dealing with "natural" ambiguities or when decisive knowledge is not yet available

*© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 17-27*

*Received 08/03/09
Accepted 27/04/09
Final version 06/05/09*

for the concerned stage of the analysis process, but most of them corresponds to incorrect interpretations.

To obtain valid final interpretations, it is necessary to control the respective relevance of the generated results using specific strategies. The goal of control strategies is to favour the most relevant results among all generated ones or symmetrically filter incorrect results.

Relying on theoretical works about knowledge base systems' control [2], [3, p. 26-43] has shown that the control of a NLP system can be considered as a decisional process where multiple heterogeneous comparison criteria have to be aggregated. This decisional formalisation of the control has conducted to an intersection between NLP and the MultiCriteria Decision Aid (MCDA) domain and more precisely outranking approaches, which propose an efficient and adapted methodology. Thus, a module dedicated to the results control based on an outranking approach has been developed and integrated as a central element of TiLT [3, p. 69-88].

This article is focused on the application of this outranking control strategy on the SMS transcription process.

Section 2 introduces the SMS transcription process and the problem of concurrent results generation. Section 3 presents the outranking control approach proposed by [3, p. 69-88] and the underlying module of control that has been integrated in TiLT's architecture. Section 4 describes the control strategy that has been defined for this particular case of SMS transcription and Section 5 gives an evaluation of this approach.

2 SMS Transcription

2.1 Related Work

SMS transcription or translation is still a recent problem and little work has been done on this topic. [4] uses a phrase-based statistical model to normalize SMS and then to translate English SMS in standard English. This task is sometimes compared to noisy text processing [5] but this approach does not take into account the particular aspects encountered in SMS. Commercial on-line software exists for French SMS translation (http://www.traducteur-sms.com or http://www.aidoforum.com/traducteur-sms.php), but it only proposes a rudimentary recognition of the most common SMS abbreviations without any linguistic processing.

[6] and [7] constitute the main references to French SMS transcription system. Both use a manually transcripted corpus of SMS proposed by the university of Louvain [8] to learn statistical language models.

Considering that such corpora were not available when we started working on SMS transcription, statistical methods were not conceivable and this is why we favoured a symbolic approach. Moreover, this applicative context constituted an interesting evaluation task for the TiLT toolbox.

2.2 TiLT for SMS Transcription

As shown in Fig. 1, the transcription process proposed by TiLT relies on the successive application of different analysis modules. First, the initial message is segmented using classical segmentation rules and specific ones like for smileys recognition. A French lexicon composed of 100 000 units enriched with 2 000 specific abbreviations (lol, msg, 2min, etc.) is used to lexically analyse each identified segment. Unknown forms are submitted to various correction and deduction strategies (typographic, morphologic, phonetic, etc.). The segments analysis generates a lattice of lexical units. This lattice is then submitted to the shallow parsing module, which regroups lexical units into chunks, and gives these chunks a syntactic label. This syntactic analysis makes use of grammatical rules, which specify constraints to be applied between chunks and internally between the lexical units which compose a chunk. Thus, the final transcription of the initial SMS corresponds to the succession of forms that has been syntactically validated.

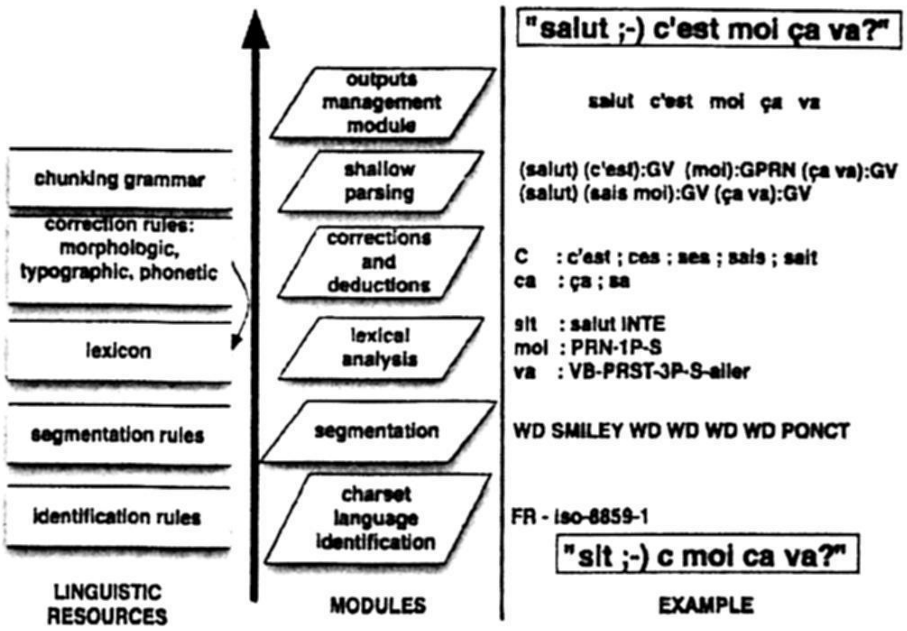
"salut ;-) c'est moi ça va?"
outputs management module
salut c'est moi ça va
chunking grammar
shallow parsing
(salut) (c'est):GV (moi):GPRN (ça va):GV
(salut) (sais moi):GV (ça va):GV
correction rules: morphologic, typographic, phonetic
corrections and deductions
C : c'est ; ces ; ses ; sais ; sait
ca : ça ; sa
lexicon
lexical analysis
slt : salut INTE
moi : PRN-1P-S
va : VB-PRST-3P-S-aller
segmentation rules
segmentation
WD SMILEY WD WD WD WD PONCT
identification rules
charset language identification
FR - iso-8859-1
"slt ;-) c moi ca va?"
LINGUISTIC RESOURCES
MODULES
EXAMPLE

Fig. 1. The initial symbolic transcription process

2.3 Indetermination

As most NLP systems, TiLT is faced with the problem of concurrent and erroneous results generation. This problematic phenomenon is even more present in such a context of spontaneous and atypical text processing.

In the example illustrated in Fig. 1, one can easily notice that most of the analysis modules that compose the transcription process are faced with concurrent and erroneous results. Indeed, the lexical analysis of ambiguous or ill-formed segments leads to concurrent lexical units. For example "C" can be phonetically corrected to "c'est", "ces", "ses", "sais", "sait", ..., or even considered as the initial of a first name like "Céline", "Cécile", ... Considering these indeterminations, the lexical analysis module generates a lattice of lexical units. Despite the fact that the shallow parsing module aims at reducing the space of concurrent lexical units using syntactic constraints, indeterminations of syntactic groups and labelling remain. Thus, to generate a transcription for an initial SMS, a selection of the one best succession of lexical units has to be performed among the remaining concurrent lexical units.

Based on a first evaluation of this initial transcription process which concerns 9 000 messages coming from the corpus of Louvain [8], we have identified and quantified the different indetermination sources.

For one input segment, the lexical analysis module and its correction strategies generates an average of 15 concurrent lexical units. These lexical units can be factorized into 3 different morpho-syntactic categories. Considering one syntactic chunking, the shallow parser produces 2.5 concurrent syntactic labellings. Finally, the succesion of forms which composes the transcription result is chosen among 2.7 concurrent lexical units for each initial segment.

Faced with these indeterminations, strong heuristics have been integrated into the initial transcription system in order to select one best final transcription. The first one concerns the selection of one syntactic chunking. As recommanded by [9], the chunking having the largest chunks is preferred. The second one is materialized by a score which is associated to some lexical or morpho-syntactic features. Defined by experts, these scores are used to select the one best syntactic labelling, the one which regroups the lexical units having the highest scores, and is also used to perform a final selection of the syntactically validated lexical units which then compose the final transcription.

The first evaluation has also emphasized the fact that 25% of the SMS are not correctly transcribed although they are completely lexically covered.

Through an analysis of these errors [10], we have noticed that for 30% of these lexically covered SMS, the syntactic chunking is wrong. For about 35% of these SMS, the preferred syntactic labelling is not completely correct. Other errors are caused by inappropriate final selection of lexical units.

3 Controlling Analysis Processes

3.1 Related Works

Control strategies aim at identifying correct interpretations among all generated ones. Obviously, this objective can only be reached if distinctive information is available to evaluate the relative relevance of the concurrent results. Thus, a control strategy first relies on the integration or declaration of comparison criteria.

This problem has largely been addressed for the ranking (or reranking) of results generated by speech recognition systems [11]. Such strategies, mainly based on empirical knowledge, have also been applied to control the results of syntactic parser [12] [13], machine translation systems [14] or even natural language generation systems [15].

The use of additional and specific knowledge to evaluate the relative relevance of concurrent results has also been investigated for more specific NLP tasks like word sense disambiguation [16] [17], machine translation [14] or coreference resolution [18].

Nevertheless, it appears that control strategies have only concerned specific applicative contexts and there is no generic formalization or methodology for controlling a complete NLP system like TiLT.

3.2 A Decisional Approach of Control

[3] has proposed to consider this task as a decision process. Based on this formalization a generic control module has been implemented and in TiLT. Indeed, as decision problems [19], a control strategy relies on a first stage of concurrent results evaluation which is then use to identify the most preferred results.

Faced with the heterogeneity of the indetermination cases, it appears necessary to combine multiple criteria during the evaluation of the results. Moreover, contrary to most of the existing control strategies which rely on statistical methods, the approach proposed by [3] makes use of expert preferences in order to determine how the concurrent results have to be compared. This way, this approach can be applied when no representative corpus is available. This formalization has led to create an intersection between NLP and a domain specialized in the resolution of such problem: MultiCriteria Decision Aid (MCDA) and more precisely outranking approaches [20] which propose methods for aggregating incommensurable criteria.

3.3 A Generic Framework of Control based on Outranking

Let $R : \{r_1, r_2, ..., r_m\}$ be the set of concurrent results and $C : \{C_1, C_2, ..., C_n\}$ the considered comparison criteria. Each criteria constitutes an increasing function which is used to evaluate the results relevance. Thus, each result $r_i, i = 1..n$

is associated to a performances vector $\{g_1(r_i), g_2(r_i), ..., g_n(r_i)\}$ which represents its evaluation on each considered criterion. As it has been previously said, the main characteristic of outranking approaches is to rely on expert preferences which determine the way the results have to be compared according to their associated performances vectors. Such preferences express the importance and the uncertainty to grant to each criteria, and also modeled incompatibility situations when two results are compared. These preferences are materialised by:

$W : \{w_1, w_2, ..., w_n\}$ a weights vector,
$Q : \{q_1, q_2, ..., q_n\}$ indifference thresholds,
$P : \{p_1, p_2, ..., p_n\}$ preference thresholds,
$V : \{v_1, v_2, ..., v_n\}$ veto/incomparability thresholds,

Q and P express an imprecision margin when two performances are compared and V define incomparability limits.

The evaluation of the concurrent results relies on a pairwise comparison in order to establish outranking relations. A result r_1 outranks a result r_2, noted $r_1 \; S \; r_2$, if a sufficient majority of criteria validates the assertion of outranking (concordance measure $c(r_1, r_2)$) and if the minority that invalidates this assertion (discordance measure $d_k(r_1, r_2), k = 1..n$) is not too strong. The concordance measure $c(r_1, r_2)$ is based on partial concordance indices $c_k(r_1, r_2), k = 1..n$ computed for each criterion:

$$c_k(r_1, r_2) = \begin{cases} 0, & \text{if } g_k(r_2) - p_k \times g_k(r_1) \geq g_k(r_1) \\ ]0,1[, & \text{if } g_k(r_1) * (1 + q_k) \leq g_k(r_2) - g_k(r_1) \leq g_k(r_1) * (1 + p_k) \\ 1, & \text{if } g_k(r_2) - q_k \times g_k(r_1) \leq g_k(r_1) \end{cases}$$

The concordance measure regroups partial concordance indices:

$$c(r_i, r_j) = \frac{1}{P} . \sum_{k=1}^{n} w_k . c_k(r_1, r_2)$$

where $P = \sum_{k=1}^{n} w_k$

The discordance is represented by partial discordance indices $d_k(r_1, r_2), k = 1..n$:

$$d_k(r_1, r_2) = \begin{cases} 1, & \text{if } g_k(r_2) - v_k \times g_k(r_1) \geq g_k(r_1) \\ ]0,1[& \text{if } g_k(r_1) * (1 + p_k) < g_k(r_2) - g_k(r_1) < g_k(r_1) * (1 + v_k) \\ 0, & \text{if } g_k(r_2) - p_k \times g_k(r_1) \leq g_k(r_1) \end{cases}$$

A global credibility index $\sigma(r_1, r_2) \in [0, 1]$ is computed from $c(r_1, r_2)$ and $d_k(r_1, r_2), k = 1..n$ and repesents the credibility to grant to the outranking relation established between r_1 and r_2.

$$\sigma(r_1, r_2) = C(r_1, r_2) \prod_{k \in \overline{F}} \frac{1 - d_k(r_1, r_2)}{1 - C(r_1, r_2)}$$

The outranking relations established between pairs of concurrent results can then be interpreted in order to produce decision recommendation of three kinds:

ranking favoring results that outrank the largest number of other concurrent results with the highest credibility degrees, a partial pre-order can be computed to represent a ranking of the concurrent results,
selection results that outrank the largest number of other concurrent results with the highest credibility degrees without being outranked by other results constitute a set of favored results
classification compared with acceptability profiles which are associated to classes by experts, results can be affected to ordered classes of equivalence.

We suggest the interested reader to read [20] and [21] for more information about algorithms used to build these decision recommendations.

These recommendations are then interpreted by analysis modules in order to favor the most preferred results or to symmetrically filter the less relevant ones.

4 Controling the SMS Transcription Process

4.1 Criteria

Based on a manual analysis of the errors made by the initial transcription process, we have remarked that many recurrent and typical SMS patterns are not well transcripted, for example: "c bon" → "c'est bon", "a plus" → " plus", "comen sa va" → "comment ca va", ...

Through a manual analysis of the erroneous transcription generated by the initial process, we have noticed that frequent and simple words successions are not correctly transcripted.

So to improve this initial transcription process, especially for such recurrent lexical and syntactic patterns, we have integrated empirical criteria in order to favor the most frequent forms and successions of form. Thus, 20 000 SMS transcriptions of the Louvain corpus have been used to establish a frequency table of lemmatized and inflected forms.

This table is first used to associate to each candidate lexical unit its observed frequency and secondly, according to this frequencies table, the Viterbi algorithm [22] is applied on the lattice of concurrent lexical units to identify one best path of words bigrams.

Therefore, the initial heuristic criterion corresponding to an *a priori* definied quantitative preference for some morpho-syntactic categories is completed with the empirical criteria. Thus, each candidate lexical unit $r_i, i = 1..m$ is then evaluated on 4 criteria:

$g_1(r_i)$ preference score on the morpho-syntactic category of r_i

$g_2(r_i)$ the frequency of r_i's lemmatised form
$g_3(r_i)$ the frequency of r_i's inflected form
$g_4(r_i)$ a boolean criteria that is true if r_i belongs to the best word bigrams path

4.2 Control Strategy

Despite the fact that the previously enumerated criteria are associated to each generated lexical unit during the lexical analysis, it appears to be inefficient to set up a control stage directly at this stage of the transcription process. Indeed, the shallow parsing aims at reducing the size of the lexical lattice throught a validation of syntactic constraints. Thus, these criteria have been first used to select the one best syntactic labelling of syntactically validated lexical units, and then to select the one best final sequence of lexical units in order to establish the final transcription.

For these two control strategies, a preferences model (Sec. 3.3) favoring empirical criteria has been defined as illustrated by table 1:

Table 1. Preferences model

criteria	weight	ind. thresh.	pref. thresh.	veto thresh.
g_1	0.3	0.4	0.6	-
g_2	0.2	0.05	0.1	0.4
g_3	0.2	0.05	0.1	0.4
g_4	0.4	-	-	-

According to the performances vectors associated to the concurrent lexical items and to this preferences model, concurrent syntactic labelling have been ranked in order to identify the most preferred one. Considering this preferred syntactic labelling and the fact that lexical indeterminations can remain for each morpho-syntactic category, concurrent final lexical units are also ranked in order to determine for each category its most preferred lexical unit and so to establish the final transcription.

4.3 Example

Let us consider the message "si tu revil j anul tt" where the form "tt" can be corrected to {"tout", "tôt","toit","tête","tant",...}. Performances vectors associated to these concurrent lexical units are illustrated in table 2 and the preferences model of table 1 is used to compare these alternatives:

One can remark that the criteria concerning form frequencies and the belonging to the best words bigram path are concordant with the assertion "tout" S "tôt" and no criterion is discordant with it. So σ("tout", "tôt") $= 0.8$ and σ("tout", "tant") = 0.8 too. Moreover, as σ("tôt", "tant") $= 0.4$, these concurrent forms are ranked in the following descending preference order: "tout" $\succ$ "tôt" $\succ$ "tant".

Table 2. Performances vectors

form	$c_1(r_i)$	$c_2(r_i)$	$c_3(r_i)$	$c_4(r_i)$
tout	12	54	54	1
tt	15	23	23	0
tant	15	18	18	0
...	...	...	...	...

5 A First Attempt of Evaluation

During a first evaluation on 9 000 messages, the initial transcription process has obtained the results presented in Table 3 using the Jaccard and BLEU measures:

$$\text{Jaccard coeff.} = \frac{|R \bigcap S|}{|R \bigcup S|}$$

where R is the set of the forms proposed by the transcription process and S is the set of forms of the solution.

$$BLEU = BP.exp\frac{1}{N}\sum_{n=1}^{N} log(\frac{\text{nb. common n-grams}}{\text{nb. n-grams}})$$

where BP is a penalty which is imposed when forms are deleted from the initial message, $BP = min(1, exp(1 - \frac{nb.wordinsolution}{nb.wordsinhypothesis})$.
nb. n-grams = (message size$-N-1$) where N is the size of the largest considered n-gram.

Table 3. Evaluation of the initial transcription process

Jaccard	BLEU	nb. erroneous forms
0.745	0.712	31 248

Table 4 illustrates the results obtained with the controlled transcription process. The improvement that seems low at first glance has to be put into perspective with the identified progression margin. Indeed, a control strategy is efficient only if at least one of the concurrent results is correct. During the first evaluation, we have noticed that only 25% of the 9 000 messages were completly lexically covered and not well transcripted. So, considering this progression gap, the control strategy has lead to a diminution of 20% of the number of final erroneous forms.

Obviously, it would have been interesting to compare our symbolic transcription system with the statistical systems of [6] and [7]. Unfortunately, we do not know on which part of the corpus [6]'s system has been evaluated. Moreover,

Table 4. Evaluation of the controlled transcription process

Jaccard	BLEU	nb. erroneous forms
0.795	0.746	29 759

they use the word error rate as the evaluation measure which is a relevant measure for statistical language model but not for our symbolic system. Indeed, this measure takes into account the number of deleted and inserted words due to the (vocal or text) signal segmentation. As our system do not perform other segmentation than the one established by whitespace characters, no words are deleted nor inserted. [7] do not propose any quantitative evaluation.

6 Conclusion and Perspectives

To overcome the problem of concurrent and erroneous results generated by the different analysis modules which compose a symbolic SMS transcription process, we have proposed a control strategy which relies on an outranking approach. Such an outranking control method allows for the aggregation of heterogeneous (empirical and heuristic) criteria which are incommensurable. The evaluation of this control strategy on 9 000 messages has shown that 20% of the erroneous and lexically covered forms are well corrected.

For future work, we are experiencing the integration of a fifth criterion corresponding to the identification of one best trigrams path of morpho-syntactic categories in the lattice of lexical units. We think that this criterion will help for the identification of recurrent syntactic patterns.

Moreover, to improve the lexical coverage of our lexical analysis module, the correction and deduction strategies have to be reconsidered. Currently, only unknown forms, forms not present in our lexicon, are submitted to deduction and correction strategies. Obviously, a lot of forms in SMS messages are mispelled but correspond all the same to known forms. In the following example "il son fou", the form "son" should be written "sont", but as "son" correspond to a valid possessive pronoun, no correction and deduction strategies are applied and no correct lexical unit is present in the lattice of concurrent forms. Symmetrically, the improvement of the lexical coverage based on the systematic use of correction and deduction strategies will induce an important decrease of the system precision but will again justify the need for control strategies.

References

1. Guimier De Neef, E., Boualem, M., Chardenon, C., Filoche, P., Vinesse, J.: Natural language processing software tools and linguistic data developed by france télécom r&d. In: Indo European Conference on Multilingual Technologies (IECMT). (2002)

2. Bachimont, B.: Le contrôle dans les systèmes à base de connaissances. HERMES (1992)
3. Smits, G.: Une approche par surclassement pour le contrle d'un processus d'analyse linguistique. PhD thesis, Universit de Caen, Orange Labs (2008)
4. Aw, A., Zhang, M., Xiao, J., Su, J.: A phrase-based statistical model for sms text normalization. In: Proceedings of the COLING/ACL on Main conference poster sessions, Morristown, NJ, USA, Association for Computational Linguistics (2006) 33–40
5. Clark, A., Tim, I.: Pre-processing very noisy text. In: Proc. of Workshop on Shallow Processing of Large Corpora. (2003)
6. C.Kobus, Yvon, F., Damnati, G.: Transcrire les sms comme on reconnat la parole. In: Actes de la Confrence sur le Traitement Automatique des. (2008)
7. Beaufort, R., Roekhaut, S., Fairon, C.: Dfinition dun systme dalignement sms/franais standard laide dun filtre de composition. In: 9es Journes internationales dAnalyse statistique des Donnes Textuelles. (2008)
8. Fairon, C., Paumier, S.: A translated corpus of 30,000 French SMS. In: Proceedings of LREC2006. (2006)
9. Abney, S.: Parsing by chunks. Kluwer Academic (1991)
10. Guimier De Neef, E., A., Park, J.: TiLT correcteur de SMS : évaluation et bilan qualitatif. In: Actes de la conférence TALN. (2007)
11. Pusateri, E., Thong, J.V.: N-best List Generation using Word and Phoneme Recognition Fusion. In: Proceedings of the European Conference on Speech. (2001)
12. Collins, M., Koo, T.: Discriminative reranking for natural language parsing. Computational Linguistics (2005)
13. Charniak, E.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the ACL. (2005)
14. Shen, L., Sarkar, A.. Och, F.: Discriminative reranking for machine translation. In: Proceedings of the Joint HLT and NAACL Conference. (2004)
15. Paiva, D., Evans, R.: Empirically-based Control of Natural Language Generation. In: Proceedings of the 43rd Annual Meeting of the ACL. (2005)
16. Rosso, P., Masulli, F., Buscaldi, D.: Word sense disambiguation combining conceptual distance, frequency and gloss. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering. (2003)
17. Dang, H., Palmer, M.: Combining Contextual Features for Word Sens Disambiguation. In: Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation. (2002)
18. Weissenbacher, D., Nazarenko, N.: A bayesian approach combining surface clues and linguistic knowledge : Application to the anaphora resolution problem. In: Proceedings of the Recent Advances in Natural Language Processing (RANLP'07). (2007)
19. Hansson, S.: Decision Theory: A Brief Introduction (1994)
20. Bouyssou, D.: Outranking approach. Encyclopedia of optimization (2001)
21. Mousseau, V., Slowinski, R., Zielniewitcz, P.: Electre Tri 2.0 Methodological guide and user's manual. Technical report, LAMSADE, Paris Dauphine (1999)
22. Forney, G.: The Viterbi algorithm. In: Proceedings of the IEEE. Volume 61. (1973)