

L'utilisation de scores numériques en sémantique computationnelle

Frédéric Landragin

LORIA – UMR 7503

Campus scientifique – BP 239

54506 Vandœuvre-lès-Nancy Cedex

Frederic.Landragin@loria.fr

1 Introduction

Dans le cadre du traitement automatique du langage naturel et plus particulièrement de la résolution ou de la production d'anaphores pronominales, de références ou encore de comparaisons et de métaphores, nous proposons une synthèse sur les méthodes de calcul utilisées pour confronter les différents facteurs de saillance privilégiant une entité du discours par rapport aux autres. Ces méthodes sont séparées en deux groupes selon que l'importance relative des facteurs est déterminée *a priori* (méthodes statiques) ou en cours de traitement (méthodes dynamiques). L'ordre de présentation va de la méthode la plus facile à la plus difficile à mettre en œuvre. Les avantages et inconvénients sont discutés d'un point de vue théorique et d'un point de vue pratique, en particulier lorsque le recours à de coûteuses analyses de corpus s'avère indispensable.

2 Les méthodes statiques

La somme ou la moyenne des facteurs. A partir du moment où l'on dispose d'une liste de facteurs de saillance (accentuation, position initiale dans l'énoncé, thème, etc., mais notre but n'est pas ici de décrire une telle liste), la méthode la plus simple consiste à identifier pour chacune des entités du discours quels facteurs privilégient sa saillance, puis de compter ces facteurs, en divisant éventuellement ensuite le total par le nombre de facteurs. Il s'agit de la moyenne arithmétique classique, qui privilégie l'entité caractérisée par le plus grand nombre de facteurs de saillance jouant en sa faveur. Cette méthode ne nécessite qu'une liste non hiérarchisée de facteurs et s'avère ainsi la plus facile à mettre en œuvre. Elle peut néanmoins détourner la théorie initiale dans le sens qu'utiliser un poids identique pour tous les facteurs induit (implicitement) que tous les facteurs de la liste initiale sont considérés comme ayant exactement la même importance. Or cette présupposition n'est souvent pas portée par la théorie initiale. Il s'agit donc clairement d'un biais qui peut rendre fautive la modélisation par quantification (sauf si la classification est présentée comme regroupant des facteurs homogènes, auquel cas la quantification reste théoriquement plausible).

En s'inspirant de la théorie de Tversky qu'ils décrivent, Iwayama *et al.* (1990) puis Pattabhiraman (1993) utilisent pour leur part la moyenne géométrique, plus précisément la multiplication de deux scores, l'intérêt étant l'influence relative des deux termes du produit. D'une manière générale, la méthode de la moyenne présente plusieurs inconvénients :

- les facteurs ont tous la même importance, or il se peut au contraire qu'un facteur ait beaucoup plus de poids qu'un autre ;
- en s'appliquant, un facteur peut en annuler un ou plusieurs autres (pénalisant d'autant l'entité auquel il s'applique) ;
- il se peut que plusieurs facteurs soient fréquemment conjoints, et que leur prise en compte incrémentale favorise une entité plus qu'il ne l'est souhaitable.

La prise en compte du facteur optimal. Il s'agit ici de classer *a priori* les facteurs de saillance par ordre d'importance, et de tester leur application sur chacune des entités du discours, en commençant par le facteur le plus important. Dès qu'un facteur s'applique, l'entité correspondante est considérée comme la plus saillante. Autrement dit, l'entité la plus saillante est celle qui satisfait le facteur le plus élevé (ou optimal). C'est une simplification du principe de la Théorie de l'Optimalité (Prince & Smolensky, 1993), qui, bien qu'initialement conçue pour la phonologie, constitue une métathéorie qui nous semble exploitable ici. Cette méthode nécessite d'être capable de fournir une hiérarchie des facteurs. C'est ce que font Hajičová *et al.* (1995) lorsqu'ils privilégient par exemple les éléments focalisés dans l'énoncé à ceux désignés par un groupe nominal dans sa partie topique. Le problème avec une telle échelle, et d'une manière générale dès qu'on considère une hiérarchie, c'est qu'une entité peut satisfaire le seul facteur optimal alors qu'une autre entité, par conséquent moins saillante, peut satisfaire une multitude de facteurs secondaires et constituer ainsi un candidat théoriquement plus pertinent. D'autre part, cette méthode induit que les facteurs de la classification théorique initiale sont non seulement hiérarchisés, mais également que la prise en compte d'un facteur annule l'intervention de tous les suivants. Ce n'est généralement pas présumé et, pour que la quantification soit théoriquement plausible, il faut que la théorie initiale soit explicite sur la hiérarchisation de ses facteurs.

La moyenne pondérée des facteurs. La pondération de facteurs selon leur importance et la prise en compte de l'ensemble des facteurs par une moyenne s'avère une solution aux problèmes des deux méthodes précédentes. Le principe est celui de la première méthode, les 1 étant remplacés par un coefficient correspondant au poids du facteur considéré par rapport aux autres facteurs. Déterminer des poids s'avère cependant délicat : l'intuition ne suffit pas à justifier des chiffres tels que 0.8 ou 0.6, et une analyse de corpus peut aboutir à des résultats biaisés de par la nature du corpus ou les difficultés que pose l'identification par l'annotateur des causes de saillance. Parmi les listes de facteurs de saillance incluant des pondérations, celle proposée par Alshawi (1987) est l'une des premières et des plus intéressantes. Si les poids sont critiquables, avec par exemple une trop grande importance donnée à la récence, l'exploitation qu'en fait Alshawi sur un large éventail de phénomènes linguistiques est en revanche appréciable. La méthode de la moyenne pondérée nous semble adéquate au calcul de la saillance mais nécessite un énorme travail de détermination des poids des facteurs. Pour être valide, ce travail devrait inclure le test systématique d'un facteur en inhibant tous les autres, puis le test de chaque paires de deux facteurs, etc. Compte tenu du nombre élevé de facteurs, la combinatoire fait que les tests nécessaires sont impossibles à réaliser.

En ce qui concerne la plausibilité théorique de cette méthode, tout dépend de la classification de départ : si elle comprend des poids, il est évident que la quantification pondérée s'en déduit naturellement et reste plausible ; si elle ne comprend pas de poids, alors les poids qui sont introduits par la quantification peuvent impliquer un biais important. Le problème qui se pose est le suivant : quelles méthodes de détermination de poids conservent une certaine plausibilité théorique, et quelles méthodes quantificatrices détruisent la théorie ? *A priori*, les méthodes intuitives ne peuvent pas démontrer leur plausibilité, et les méthodes de détermination de poids à partir d'analyses statistiques de corpus ne restent plausibles que dans le cadre d'un corpus ou d'un type de corpus donné (textes scientifiques, textes journalistiques, etc.). Pour que les poids soient génériques, c'est-à-dire indépendants du corpus d'étude, il faut qu'ils soient significativement équivalents entre différents corpus testés en nombre suffisant. Cependant, même dans le cas de la restriction à un corpus particulier et à la quantification de la théorie dans ce seul corpus, rien ne dit que les poids sont valides : tout dépend de l'interprétation des phénomènes que font les annotateurs. Dans l'exemple de la quantification de la saillance, il est déjà extrêmement difficile d'identifier tous les facteurs qui sont intervenus pour rendre une entité du discours saillante, et il est encore plus difficile de dire

quel facteur a été prépondérant (du moins quand les facteurs ne se limitent pas à des fonctions grammaticales non ambiguës). Il est ainsi nécessaire de faire intervenir plusieurs annotateurs, et de recourir à des méthodes de comparaison des résultats – par exemple celle de l'indice kappa détaillée dans (Carletta, 1996) – qui peuvent s'avérer très lourdes à mettre en œuvre.

Les méthodes procédurales. Pour résoudre des anaphores pronominales, Mitkov (1998) définit des heuristiques basées sur des scores intuitifs. Son approche caractérisée par le peu d'information exploitée ne nécessite même plus une analyse syntaxique complète. Les facteurs de saillance en sont très réduits, comme le montre la liste suivante où l'opération sur le score de saillance est indiquée entre parenthèses : détermination (0 pour un défini et -1 pour un indéfini) ; information connue (1 pour le premier groupe nominal qui est assimilé au thème, 0 sinon) ; nature du verbe (le score varie selon que le verbe appartient à une liste prédéfinie) ; répétition lexicale (score de 0 à 2 selon le nombre de répétition) ; etc. Cette approche illustre le recours à une méthode calculatoire extrême, avec son avantage d'être opérationnelle et parfois pertinente, et ses nombreux inconvénients. Ceux-ci sont énumérés par Salmon-Alt (2001), qui souligne en particulier que les choix ne sont absolument pas plausibles d'un point de vue linguistique : « pourquoi un indéfini a-t-il moins de chances d'être l'antécédent d'un pronom ? D'où vient la liste des verbes et des connecteurs ? Comment justifier la pondération des scores ? ».

Même si le but de Mitkov est d'aboutir à une modélisation numérique sur une tâche très restreinte et en aucun cas d'expliquer ni même de s'intégrer dans une théorie linguistique, la question se pose de la plausibilité théorique des méthodes procédurales. Soit elles cherchent à traduire en procédures des règles linguistiques et elles restent alors *a priori* théoriquement plausibles, soit elles cherchent à optimiser un calcul en construisant elles-mêmes des procédures qui ne découlent d'aucune théorie, et dans ce cas elles s'avèrent clairement éloignées de toute théorie et de toute validité. En ce qui concerne la problématique de la saillance, le premier cas est inexistant : il faudrait que la linguistique et la pragmatique élaborent une théorie à la Chomsky avec des règles précises qui articulent les éléments de la structure informationnelle. Quant au deuxième cas, l'exemple typique de Mitkov pose la question de l'intérêt de telles méthodes : les procédures ne sont pas absolument pas plausibles d'un point de vue linguistique, et pourtant elles permettent la réalisation de la tâche, c'est-à-dire une résolution satisfaisante des anaphores pronominales. Tout ce qu'on peut en déduire, c'est qu'il existe un moyen informatique de résoudre une tâche restreinte, avec des résultats comparables au fonctionnement humain. En aucun cas on ne peut affirmer que ce moyen a quelque chose à apprendre à la linguistique ou au fonctionnement cognitif humain. Le jour où on met en œuvre un moyen informatique plus puissant, le moyen utilisé précédemment est tout simplement jeté. L'intérêt des méthodes procédurales s'avère ainsi bien faible.

Les méthodes statistiques et les approches hybrides. Pattabhiraman (1993) utilise un réseau de relations statistiques entre concepts pour identifier la catégorie la plus saillante dans une situation donnée. Les résultats s'avèrent convaincants et montrent l'intérêt de certaines méthodes statistiques. C'est le cas de l'analyse factorielle des correspondances permettant (à condition de disposer de l'avis d'experts) de déterminer les influences relatives des divers facteurs dont nous avons parlé, par exemple l'influence de la position initiale et de la fonction sujet sur le statut de thème. Cette idée reste cependant à l'état de perspective de recherche. Une autre perspective qui nous semble intéressante dans ce domaine est l'exploration de méthodes hybrides, telle que la combinaison d'une méthode basée sur la moyenne pondérée de facteurs avec une méthode statistique, celle-ci remettant en question les différents poids compte tenu des interférences entre facteurs.

3 Les méthodes dynamiques

Figurer une hiérarchie de facteurs peut sembler dangereux : rien ne dit que dans un contexte ou un autre, tel facteur prendra une importance toute particulière. Exemples : dans une suite de deux phrases où le thème de la seconde reprend le rhème de la première, le facteur lié à la distinction entre thème et rhème n'a pas le même poids dans les deux phrases ; lorsque le propos est identifié et considéré comme saillant, ce calcul doit rester activé lors de l'analyse des phrases à venir. Le mieux serait de gérer dynamiquement une hiérarchie des facteurs de saillance, pour que celle-ci ne soit pas trop déterminante.

Un autre exemple où l'importance des facteurs de saillance est gérée dynamiquement est celui de Lappin & Leass (1994), qui se basent sur le travail d'Alshawi (1993) pour proposer un algorithme de résolution des anaphores pronominales. Leur algorithme ne consiste pas réellement en une moyenne : il part d'un seuil initial qui varie au fur et à mesure du traitement de phrases. Certaines étapes consistent par exemple à diviser par 2 le poids de tel facteur, d'autres consistent en l'application de filtres (morphologiques ou syntaxiques). Même si cette approche présente quelques défauts (récence privilégiée, limitation à des facteurs purement formels) et reste à améliorer, elle constitue une première étape dans la gestion dynamique de scores qui nous semble théoriquement intéressante. Cette méthode est cependant très difficile à mettre en œuvre (il s'agit d'identifier l'ensemble des influences contextuelles sur chacun des facteurs de saillance) et, face aux difficultés, les auteurs avouent eux-mêmes que les poids qu'ils proposent sont arbitraires : « the specific values of the weights are arbitrary ».

4 Discussion

Le principal constat que nous pouvons faire à ce stade est ainsi le suivant : si les méthodologies mathématiques et statistiques semblent stabilisées, il en n'est pas de même de celles relatives à la détermination des poids et des influences relatives des divers facteurs et éléments contextuels, et, pour ce faire, des méthodologies relatives au traitement de corpus. C'est dans ce sens qu'il nous semble important de continuer les recherches, avec les préoccupations suivantes :

1. Dans tout travail de quantification, il faut avoir en amont des hypothèses à vérifier. Sans vision pré-théorique, on ne sait pas où on va, on ne sait pas ce qu'on veut montrer, et les quantifications n'ont pas de sens. La difficulté, mais aussi tout l'intérêt de la démarche, réside dans la formulation d'une traduction d'hypothèses en données, par exemple sous la forme d'un schéma d'annotation.

2. Dans tout travail de quantification, il faut avoir en aval des conclusions à tirer sur le comportement du système. Autrement dit, la quantification ne doit pas uniquement servir à proposer une solution numérique à un problème donné, ne doit pas servir qu'elle-même. Elle doit également permettre de renforcer voire de valider la théorie de départ. Elle peut ainsi servir à identifier des situations où une analyse plus fine est nécessaire, ou encore à expliquer certains phénomènes. Contrairement à Mitkov qui n'explique rien du fait de sa tâche très réduite, il nous semble important d'avoir un but plus ambitieux – l'interprétation, l'accès au sens – et d'exploiter les méthodes de quantification comme un tremplin vers ce but, comme une étape intermédiaire et provisoire dont l'objectif est de stabiliser un état de connaissances.

3. Toute étude de corpus doit être faite avec le plus grand soin. Si de nombreuses mesures et de nombreux indices comme l'indice kappa sont actuellement utilisés, il reste encore à caractériser de manière précise un corpus. Or cette tâche s'avère extrêmement vaste : un corpus peut être très faible au niveau lexical et pourtant très riche au niveau référentiel. Il faut ainsi caractériser un corpus selon tous les plans considérés. Dans le cas de la quantification de

la saillance, ces plans sont ceux de la prosodie, du lexique, de la morphologie, de la syntaxe, de la sémantique, de la pragmatique... quasiment toutes les facettes de la linguistique !

Références

- Alshawi, H. (1987). *Memory and Context for Language Interpretation*. Cambridge : Cambridge University Press.
- Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics* 22(2), pp. 249-254.
- Hajičová, E., Hoskovec, T. & Sgall, P. (1995). Discourse Modelling Based on Hierarchy of Saliency. *Prague Bulletin of Mathematical Linguistics* 64, pp. 5-24.
- Iwayama, M., Tokunaga, T. & Tanaka, H. (1990). A Method for Calculating the Measure of Saliency in Understanding Metaphors. *Proceedings of the 8th National Conference on Artificial Intelligence (AAAI'90)*, Boston, pp. 298-303.
- Lappin, S. & Leass, H. (1994). A Syntactically Based Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics* 20:4, pp. 535-561.
- Mitkov, R. (1998). Robust Pronoun Resolution with Limited Knowledge. *Proceedings of the 18th International Conference on Computational Linguistics*, Montréal.
- Pattabhiraman, T. (1993). Aspects of Saliency in Natural Language Generation. Ph.D. Thesis, Simon Fraser University.
- Prince, A. & Smolensky, P. (1993). Optimality Theory: Constraint Interaction in Generative Grammar. Technical Report, Rutgers University.
- Salmon-Alt, S. (2001). Référence et dialogue finalisé : de la linguistique à un modèle opérationnel. Thèse de doctorat, Université Henri Poincaré, Nancy.