email=anca.pascu@univ-brest.fr,

Regard critique sur la « simplification d'un texte »

Anca Christine Pascu

Université de Bretagne Occidentale, Brest, France

Abstract

L'activité de « production des textes », vieille depuis des siècles ouvre un nouveau débat avec le développement des nouvelles technologies de l'informations et l'explosion de la quantité d'informations (données massive). Dans cet article, nous nous proposons une analyse du sous-domaine du traitement automatique du langage (TAL) « simplification de texte ». Nous faisons une analyse de ce qu'un texte représente par rapport aux trois dimensions de la pensée :
- Le raisonnement (la logique).
- Le langage avec ces trois propres dimensions (sous-domaines) : le lexique, la syntaxe et la sémantique.
- La relation entre les trois sous-domaines de la linguistique par rapport aux actuelles approches de « traduction assistée par l'ordinateur » (TAO).

Par rapport à cette analyse, on discute les conditions restrictives de la simplification en relation avec les enjeux positifs et négatifs induits par la simplification. On fait une analyse critique de la modélisation jusqu'à l'implémentation - machine d'un « outil de simplification de texte ». Finalement, on présente un exemple de « simplification sémantique » obtenue par deux outils de simplification.

Keywords

Texte, Simplification de texte, Traitement Automatique du Langage (TAL), ontologie.

1. Introduction

Nous analysons le domaine « la simplification de texte » comme un sous-domaine interdisciplinaire.
L'évolution de l'informatique à partir de premiers ordinateurs et premiers langages de programmation et jusqu'aux système imbriqués et aux outils de type matériel (les robots ....) ou logiciel (les langages de programmations comme Python, les bases de données intégrées ...) d'aujourd'hui s'est produit d'une manière de plus en plus multidisciplinaire avec l'apparition des sous-domaines comme prolongation des anciens domaines traditionnels : la logique quantique pour la logique, des différentes théories sémantiques pour la linguistique, la théorie mathématique sous-jacente à la théorie de l'information.
Les multiples caractéristiques interdisciplinaires de cette évolution ont été synthétisées dans sa dénomination anglaise « computer science » et plus tard en intelligence artificielle (IA) encore plus multidisciplinaire.
L'activité de « production des textes », vieille depuis des siècles suit un même developpement

Mots/Machines-2022: Journée d'étude, 25 March, 2022, Laboratoire HCTI, Université de Bretagne Occidentale, Brest, France

et, par conséquence, ouvre un nouveau débat avec le développement des nouvelles technologies de l'informations et l'explosion de la quantité d'information (données massive).
Pour mieux voir ce qui est derrière « la simplification de texte », on part de sa définition donnée par Wikipédia (4, 5)

Definition La simplification de texte (ST) (Text Simplification (TS)) est une opération utilisée en traitement automatique du langage naturel pour modifier, augmenter, classifier ou traiter autrement un corpus existant de texte lisible pour l'homme de telle manière que la grammaire et la structure de la prose soit considérablement simplifiée, tandis que la signification fondamentale et l'information restent les mêmes. La simplification de textes est un domaine de recherche important, le langage humain habituel contenant des constructions composées complexes qui ne sont pas facilement traitées automatiquement. Une approche récente consiste à simplifier automatiquement le texte en le convertissant en anglais de base; qui a un vocabulaire de seulement 1,000 mots qui sont également utilisés pour décrire dans les notes de bas de page la signification de 30 000 mots du Dictionnaire de base de la science.

Les éléments de cette définition sont :

- La simplification de texte (ST) est une opération utilisée en Traitement Automatique du Langage (TAL).
- Traiter autrement (?).
- La grammaire est simplifiée, la prose (?) est simplifiée.
- Les conditions sont : la signification fondamentale et l'information restent les mêmes.

Les questions qui se posent sur ces éléments définitoires sont :

- Qu'est-ce que « traiter autrement » exactement ?
- Que représente « simplifier la grammaire » ?
- Que représente « simplifier la prose » ?

Regardons la définition du concept « signification » (6).

«Le mot signification possède plusieurs acceptions. C'est d'abord et principalement le sens d'une chose, d'un symbole, d'un mot, ensuite l'acte de notification que l'on fait, connaissance que l'on donne d'un arrêt, d'un jugement, d'un acte, par voie judiciaire et légale, par ministère d'huissier» .

De toute cette « ambiguïté conceptuelle » il en résulte les questions suivantes :

1. Qu'est-ce qu'un texte ? De quel type parle-t-on ?
2. Que représente la simplification de la grammaire ?
3. Que représente la simplification de « la prose » ?
4. Par quoi se distingue une « opération » de simplification de texte du résumé de textes ?
5. Quel est le changement de paradigme opéré par le traitement automatique du langage (TAL) dans le traitement des textes ?
6. Quels sont les enjeux cognitifs du développement de la « simplification de texte » ?
7. Quels sont les enjeux pédagogiques du développement de la « simplification de texte » ?

Nous allons à analyser ces questions dans les sections suivantes de cet article.

2. Texte, représentation de textes, type de textes, classification de textes

Il n'est pas simple de donner une définition générale du texte parce qu'il est difficile de trouver un ensemble de caractéristiques communes à tous les textes. C'est pour cette raison qu'on pourrait dire d'une manière générale : **Un texte est une forme de représentation de connaissances.**

Il y a deux autres « traits cognitif s » propre à tout texte, « l'argumentation » et « le message ». Le message représente « ce qu'on veut transmettre cognitivement ou affectivement par ce texte ». L'argumentation représente le mode de transmission.

Tout d'abord, il faut classifier les textes dans deux catégories :

- Textes bien écrits
- Textes mal écrits

Les textes bien écrits doivent respecter les conditions de base d'un système logique :

- Cohérence.
- Consistance.
- Non-contradiction.

L'argumentation doit respecter en quelque sorte les règles d'une « inférence logique ». Par analogie avec la logique, le message doit être une sorte de conclusion de l'argumentation.
En ce qui concerne la classification de textes, qu'elle soit automatique ou manuelle elle peut se faire selon plusieurs critères. Parmi les critère de classification, nous mettons en premier le domaine des connaissances sur lequel porte le texte. Dans ce sens, on a les textes scientifiques, les textes littéraires, les textes de presses, etc. Chaque catégorie ci dessus a des sous-catégories. L'arbre des critères est défini par le contenu du domaine auquel le texte appartient. Le domaine d'appartenance du texte attribue une certaine rigueur au texte et à l'argumentation de son contenu. Il a aussi un impact fort sur la conclusion vue comme étant le message. Un autre critère très important, surtout en traduction c'est la langue du texte.
Nous affirmons que les critères ne sont pas complètement indépendants les uns des autres. Même un texte de mathématique qu'on peut considére comme le type de texte le plus « standardisé » n'est pas complètement independant de la langue dans laquelle il est écrit.

3. La simplification de texte

Dans cette section, nous analysons « la simplification de texte » comme sous-domaine du Traitement Automatique du Langage (TAL).

Les trois questions qui se posent sont :

- Pourquoi fait-on une simplification d'un texte ?
- Comment construire une approche tel que par son implémentation-machine on obtienne un texte simplifié le plus proche possible du texte initiale ?
- Quel est la différence TAL entre « résumé de texte » et « simplification de texte » ?
- Comment faire le texte "compréhensible" par une large catégorie de lecteurs (la vulgarisation) ?

3.1. Les paramètres

Pour que la simplification de texte devienne un sous domaine du Traitement Automatique du langage (TAL) au sens scientifique du terme et non pas un "bricolage" entre approches scientifiques, technologies et enjeux de temps et de ressources, il faut une analyse de son opportunité en tenant compte d'un certain nombre de paramètres.
Ces paramètres sont (la liste n'est pas exhaustive) :

1. Le public auquel le texte s'adresse : âge, éducation, catégorie sociale.
2. Le message du texte représente l'ensemble d'idée qu'on voudrait trasmettre par le texte.
3. Le type du texte scientifique, littéraire etc.
4. Le contenu du texte : le message,les concepts, l'argumentation, le degré de symbolisme.

1. 2. 3. On prend comme exemple la littérature pour les enfants, le groupe d'âge 6-8 ans et comme texte le chapitre du roman de Victor Hugo, Les Misérables où l'auteur décrit la vie de Cosette chez Les Thénardiers. Le message pour un enfant est de s'élargir son niveau de connaissances sur l'histoire et sur les réalités de la vie, d'un part et de s'enrichir le vocabulaire d'autre part. La simplification d'un tel texte est possible pour le rendre accesible à la catégorie d'âge, mais surtout pas une simplification par un logiciel de simplification.
Un autre exemple peut être le suivant : la présentation d'un texte de théorie quantique simplifié à un agriculteur par rapport à sa présentation à un professeur de physique censé de l'expliquer à ses élèves. Pour le premier, le message est neutre, tandis que pour le deuxième, le message est extrêmement important.
4. Ces éléments sont :

- Le **message** est le contenu qu'on voudrait transmettre par le texte.
- Les **concepts** sont les notions fondamentales qui apparaissent dans le texte ou, éventuellement, en dehors du texte mais en relation avec certaines notions importante du texte.
- L'**argumentation** est le chemin explicatif qui justifie le message du texte.
- Le **degré de symbolisme.** La plupart des disciplines scientifique utilisent des symboles dans la représentation de leurs théories. La suppression de certains symboles peut nuire au contenu.

Entre ces paramètres, il y a parfois des oppositions qui rendent la simplification soit inutile soit non réalisable automatiquement (par machine).

4. Exemples

Nous avons regardé deux outils de simplification textes : Smodin (7) et Paraphraz.it (8)
La définition d'un outil de simplification de textes publié sur le site web de Smodin est :

Un réécrivain, également connu sous le nom de machine de paraphrase, de réécriture de paragraphe ou de réécriture de texte, est une machine qui reformule une phrase ou un paragraphe en modifiant la séquence de mots, en utilisant d'autres mots pertinents ou en ajoutant un contexte supplémentaire. Dans certains cas, comme avec le rewriter Smodin, il peut parfois améliorer l'écriture

et la rendre plus concise.
On ne peut pas considérer cette description comme une définition. Elle n'est pas fausse, mais elle contient des ambiguités. Qu'est-ce que ça veut dire exactement "améliorer l'écriture" ? Cela porte sur la longueur du texte, sur l'accès à la compréhension....?

4.1. Un éxemple de Smodin

Voici un exemple traité par Smodin. C'est un extrait de Edward Sapir, Selected Writings in Language, Culture, and Personality, University of California Press, 1985. (9)

En ordre, il y a l'original, la traduction française et la simplification opérée par Smodin.La traduction en français est de Catherine Détrie, Paul Siblot et Bertrand Vérine, Termes et concepts pour l'analyse du discours : Une approche praxématique, Paris, Honore Champion, coll. « Bibliothèque elzévirienne », 2001 (10).

« It is quite an illusion to imagine that one adjusts to reality without the use of language and that language is merely an incidental means of solving specific problems of communications or reflection. The fact of the matter is that the 'real world' is to a large extent built up on the language of the group. No two languages are ever sufficiently similar to be considered as representing the same reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached. » (Sapir, 1949)
Il est tout à fait illusoire d'imaginer que l'on s'adapte à la réalité sans l'usage du langage et que celui-ci n'est qu'un moyen accessoire de résoudre des problèmes spécifiques de communication ou de réflexion. Le fait est que le "monde réel" est, dans une large mesure, construit sur la langue du groupe. Il n'y a jamais deux langues suffisamment semblables pour être considérées comme représentant la même réalité. Les mondes dans lesquels vivent les différentes sociétés sont des mondes distincts, et non pas simplement le même monde auquel sont attachées des étiquettes différentes.

Il est tout à fait illusoire d'imaginer une personne s'adaptant à la réalité sans utiliser le langage, qui n'est qu'une aide à la résolution d'un problème précis de communication ou de réflexion. En fait, le "monde réel" est largement basé sur le langage du groupe. Il n'y a pas deux langues suffisamment similaires pour être considérées comme représentant la même réalité. Les mondes dans lesquels vivent différentes sociétés sont des mondes différents, pas seulement le même monde avec des étiquettes différentes.

Ce qu'on pourrait remmettre en cause dans la simplification tient du type de texte. Le texte porte sur l'hypothèse de Sapir-Worf sur le rapport entre le langage et le mode réel, hypothèse connue dans la linguistique sous le nom d'"universalisme". Donc, les modifications " une personne s'adaptant" et "la résolution d'un problème" le sens est légèrement différent.

Le texte suivant : *L'astuce consiste à découvrir une relation entre deux systèmes de pensée, deux raisonnements apparemment opposés. Est-ce l'opposition, est-ce vraiment l'opposition ? C'est une*

question ouverte. C'est une des différences entre stéréotype et ingéniosité, "langage de bois" et "langage constructif", "esprit dogmatique" et "esprit scientifique".

est presque identique. Le mot "stéréotype" est remplace par "rigidité".

L'astuce consiste à découvrir une relation entre deux systèmes de pensée, deux raisonnements apparemment opposés. Est-ce l'opposition, est-ce vraiment l'opposition ? C'est une question ouverte. C'est une des différences entre "rigidité" et "ingéniosité", "langage de bois" et "langage constructif", "esprit dogmatique" et "esprit scientifique".

Ce qu'on constate est le fait que en dépit de texte qui est très court, donc il n'y a pas besoin de l'écourter, il n'y a pas une "modification sémantique". Le texte mofifié à le même sense que le texte initial.

4.2. Un éxemple de Paraphraz.it

Le même paragraphe de Sapir est modifié que dans la phrase suivante :

Les mondes dans lesquels vivent différentes sociétés sont des mondes **séparés** , pas simplement le même monde avec des étiquettes différentes.

A notre avis, la modification proposée change légèrement le sens.

5. Une proposition de modélisation – intégration dans une bonne interdisciplinarité

Pour l'instant, nous ne pouvons pas préciser la différence entre "résumé du texte" et " simplification du texte", en ce qui concerne son paradigme au sein de la communauté de l'intelligence artificielle (IA). Ce sont les modifications les caractéristiques propres à la simplification, et la longuer la caractériqtique propre au résumé ? Il faudra peut-être une définition plus précise en se tenant aux caractéristiques de texte précisées dans la section 3.1.

Notre proposition porte sur l'affinement des logiciels de simplification. Il s'agit de l'intégration d'une analyse sémantique plus profonde du texte à simplifier au logiciel (système) de simplification.

Cette approche est basée sur l'idée que chaque texte peut avoir une "représentation sémantique". La représentation sémantique doit être une représentation du texte avec deux parties:

- La structure sémantique. Elle représente l'ensemble des concepts qui font une sorte de "colonne vertébralle conceptuelle" du texte.
- L'argumentation. L'argumentation représente l'enchainement des inférences logiques qui font le discours.

La modélisation du système qui réalise une telle simplification doit suivre les étapes suivantes:

1. Le positionnement du texte dans une catégorie définie à priori selon le domaine de connaissance, le public de déstination, le type de message à transmettre, le type d'argumentation. Cette opération nécessite un grand travail interdisciplinaire et des connaissances multiples.
2. L'application d'une "annotation sémantique" pour identifier les concepts significatifs et essentiels qui définissent la structure sémantique du texte.
3. La construction d'une "ontologie du texte" qui doit contenir les concepts signiticatifs. L'ontologie d'un texte est un réseau ayant comme noeux les concepts et comme liens les relations entre eux.
4. L'extraction de deux sous-réseaux, l'un qui doit être la représentation du texte simplifié et l'autre qui doit être la représentation de la partie du texte qui porte sur le message. Le choix de ce sous-réseau du texte simplifié se faira en fonction du sous-résaux correspondant au message du texte.
5. La construction du texte simplifié par "génération de texte assistée par l'ordinateur".

6. Conclusions

Cette petite étude sur la simplification de texte conduit à quelques conclusions qui s'imposent étant surtout en rapport avec les paramètres formulés dans la section 3.1 :

1. La première et la plus importante est liée au but de la simplification. Supposons qu'on a à simplifier un texte "bien écrit". Pourquoi on a besoin de le simplifier ?

   a) Pour le lire plus vite ?
   b) Pour économiser l'espace de son support ?
   c) Pour le rendre accessible à une catégorie plus large de lecteurs.

   Dans ces deux cas ci-dessus, le résumé ferait bien l'affaire. Donc, les modifications ne sont plus utiles. Dans le troisième cas pour concevoir un bon texte proche de la vulgarisation, il est absolument nécessaire une approche sémantique de simplification.
   On peux donner au moins deux exemples qui prouvent l'utilité de la simplification :
   a) La littérature pour les enfants selon l'âge.
   b) La vulgarisation d'un texte scientifique destiné au grand public.

   Dans les deux cas ci-dessus l'analyse sémantique est très importante.
2. La simplification de la grammaire qui est revendiquée par certains outils de simplification à notre avis, n'est pas un argument pour l'utilité de la simplification.
3. L'analyse sémantique du texte doit précéder comme importance dans toute hiérarchie de paramètres. La facilité de la "compréhension du texte" ne doit pas transgresser les conditions d'un système logique (cohérence, consistence et non-contradiction). Cela confère au texte une sorte de "rationalité".
4. Pour des raisons pédagogiques (l'apprentissage de nouveaux concepts, la maîtrise de la grammaire d'une langue, une bonne traduction d'une langue à l'autre), la simplification doit rester entre certaines limites.

5. La simplification de texte en tant que sous-domaine du Traitement Automatique du Langage (TAL) reste une ”simplification assistée par l'ordinateur”.
6. La simplification de texte doit être pratiquée d'une manière intelligente, scientifique, à l'opposé d'un ”bricolage” de techniques et méthodes plus ou moins scientifiques.

7. Reférences

1. « Simplish Simplification and Summarezation Tool » [archive], The Goodwill Consortium (consulté le 28 septembre 2019).
2. Anca Christine Pascu, Modeling a software of semantic text analysis, Mots-Machines - Le sens de l'humour, Université de Brest, 5 mars, 2021.
3 Tzu-Keng Fu and Anca Christine Pascu, Conceptual Metaphor in Teaching Logic, Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education.
4. Wikipedia, Text Simplification
5 Wikipedia, Simplification de texte.
6. Wikipedia, Signification.
7. https://smodin.io/fr/reformuler-automatiquement-le-texte-en-francais-gratuitement
8. https://paraphraz.it/fr/
9. Edward Sapir, Selected Writings in Language, Culture, and Personality, University of California Press, 1985.
10. Catherine Détrie, Paul Siblot et Bertrand Vérine, Termes et concepts pour l'analyse du discours : Une approche praxématique, Paris, Honore Champion, coll. « Bibliothèque elzévirienne », 2001.