LE DICTIONNAIRE DE MACHINE RUSSE-ANGLAIS DE L'EURATOM

Perschke

Nous travaillons à Ispra sur la traduction automatique russe-anglais, et je pense qu'il y a des problèmes communs. Le problème commun le plus important est la polysémie des mots qui existe dans toutes les langues. Dans mon cas spécifique, il faut distinguer deux niveaux de polysémie; l'une se réfère seulement à la langue d'entrée, c'est-à-dire qu'un mot russe, ou d'une autre langue, peut avoir plus d'une signification. Par exemple, le mot italien "amo" peut être un verbe ou un nom. L'autre problème est que dans une langue d'entrée il y a un mot univoque, mais qui dans une langue de sortie doit être traduit par différents mots, selon le contexte. Par exemple, si nous prenons le mot anglais "to know", il est, pour l'anglais, univoque: mais si vous traduisez en italien ou en français, il faut utiliser le verbe "savoir" ou "connaître". Il faut trouver, pour tous ces cas,des règles ou des conditions,en premier lieu pour définir la signification d'un mot dans un certain contexte; en deuxième lieu, pour définir ainsi la traduction du mot correct pour la langue de sortie. Le problème de la polysémie se pose sur deux niveaux: lors de la compilation d'un dictionnaire, on se base uniquement sur la compréhension des mots; il faut avoir quelqu'un qui comprenne les mots et ce que le mot signifie. D'autre part, quand on trouve un mot dans le texte à traduire, on le cherche dans le dictionnaire et on trouve que ce mot peut avoir différentes significations ou seulement différentes traductions. Le problème est de trouver les conditions d'attribution des mots du texte à une des plusieurs valeurs de l'entrée du dictionnaire pour pouvoir sortir dans la manière juste. Nous avons trouvé que la méthode la plus efficace pour résoudre ce problème est l'analyse syntaxique, parce qu'il y a assez peu de mots qui sont équivalents, disons ambigus, dans la langue d'entrée, et qui ont la même fonction syntaxique; par exemple "amo" = verbe, et "amo" = nom, sont écrits dans la même manière, mais si on trouve les mots dans uh contexte, ils auront certainement des fonctions différentes dans la proposition, et, définissant la structure de la phrase, on pourrait très bien éliminer la possibilité d'une des significations. Je pense que pour la polysémie c'est tout ce que je peux dire.

Je pourrais dire un mot sur notre travail en général. Nous utilisons maintenant un système de traduction automatique russe-anglais, qui a été développé par l'Université de Georgetown. Nous l'avons développé et amélioré depuis 1953, et maintenant il marche à une vitesse de traduction d'environ 60000 mots par heure. Les traductions sont utilisées par les chercheurs du centre atomique qui demandent des textes d'articles scientifiques russes et ne connaissent pas le russe. La traduction elle-même n'est pas très belle, et nous avons constaté que les anglais n'apprécient pas beacoup le langage produit, parce que ce n'est pas de l'anglais correct, beau; mais il est possible, d'autre part, de comprendre ce texte et de l'utiliser pour les informations scientifiques qui ne sont pas accessibles sans traduction.

Busa

Avant tout, c'est une donnée de fait très intéressante, que vous fassiez la traduction automatique à la vitesse de 60.000 mots et que cette traduction soit employée pratiquement par les chercheurs de l'Euratom. Vous avez alors un dictionnaire de machine russe-anglais?

Perschke

Nous avons en machine un seul dictionnaire russe-anglais. Il n'y a pas une langue intermédiaire, un système intermédiaire, qui permettraient le couplage, mais chaque entrée du dictionnaire a comme argument de recherche le mot russe, l'information grammaticale et syntaxique nécessaire; il n'y a pas beaucoup d'information sémantique.

Busa

Et dans ce dictionnaire est-ce-que les mots sont groupés selon constantes, avec les désinences sur une autre bande ou bien les mots entiers sont-ils flexionnés?

Perschke

Les préfixes n'ont pas une fonction morphologique. Il y a des études sur les fonctions des préfixes, mais une application pratique doit être sûre. Nous avons vu que le détachement des préfixes en plusieurs cas est très dangereux. Aussi pour plusieurs cas il est possible de dire que "pro" ou "pri" sont des préfixes. Dans le système actuel il y a un détachement des préfixes, mais plus pour les préfixes dans la terminologie scientifique, comme "semi" "poly", etc... Il y a une certaine collection qui permet d'obtenir un dictionnaire un peu plus petit.

Busa

Pour la section anglaise du dictionnaire avez-vous aussi les "stems" et les désinences?

Perschke

Dans la forme actuelle, il y a une forme standard dans le dictionnaire, plus une indication de la manière dont cette forme standard est modifiée en dépendance de la forme morphologique, qui définit les règles de synthèse de l'anglais. Par exemple, si on a un substantif anglais terminant par "y", ce mot est codifié comme se terminant par "y" avec un code qui signifie que l'y dans un pluriel est remplacé par "ies".

Busa

Pourriez-vous nous donner le pourcentage des homographes que vous avez dans votre dictionnaire?

Perschke

Nous n'avons pas fait une étude spécialisée sur les homographes. Nous appliquons, dans la recherche du vocabulaire, le principe qui s'appelle en américain "the longest match", c'est-à-dire la correspondance la plus longue. Quand on fait l'analyse morphologique, on accepte comme la plus probable la solution qui trouve la racine la plus longue, le suffixe le plus court.

Busa

Mais avez-vous des routines pour sélectionner certains homographes sur la base du contexte?

Perschke

Il y en a un certain nombre certainement; normalement on codifie ces homographes comme forme pleine, comme individuelle, et on y attache les règles pour les distinguer.

Busa

Combien? Des centaines ou des dizaines?

Perschke

Des centaines, je pense. Le dictionnaire que nous utilisons actuellement contient à peu près 35.000 racines. Nous avons reçu des américains un dictionnaire de 180.000 entrées. Maintenant nous l'analysons et nous espérons

pouvoir l'utiliser dans le système que nous sommes en train de développer. Il sera prêt, je pense, en 1970.

Busa

Je pose une dernière question, à vous et à M. Ott de l'Université de Tübingen, nos deux collègues allemands ici: existe-t-il un dictionnaire de machine allemand? Le R.P. Valentini a précisément un problème avec l'allemand. Nous avons à Gallarate un tout petit dictionnaire allemand, encore en fiches, extrait de 50.000 mots de Kant et de 50.000 mots de Goethe. Mais je ne sais pas s'il y en a d'autres.

Perschke

L'Université de Saarbrücken a conduit des recherches sur l'allemand et a analysé deux corpus sur une collection d'articles de journaux de la Frankfurter Allgemeine Zeitung, et un autre corpus dont je ne me souviens pas bien. On a élaboré un texte de 200.000 mots courants et on a fait un dictionnaire-machine et aussi un programme pour l'analyse syntaxique de l'allemand. Il semble qu'il marche assez bien.

Busa

Connaissez-vous les travaux de M. Mater de Berlin-Est?

Perschke

Je ne les connais pas.

. . . . . .

Il est très difficile de faire une évaluation de la qualité. Je pense qu'au moins 98°/₀ des mots se trouvent dans le dictionnaire, c'est-à-dire 2°/₀ des mots ne sont pas traduits avec le dictionnaire actuel de 35.000 entrées. Pour l'analyse syntaxique et l'analyse de la polysemie, c'est très difficile. Par exemple, nous avons vu que pour la compréhension la manière dont est traduite une préposition n'est pas très importante.

D'autre part, il est très grave qu'un mot-clé du texte soit traduit d'une manière erronée. Pour cette raison, il est très difficile d'évaluer le pourcentage des erreurs ou des traductions correctes. Nous avons vu qu'il est possible de comprendre toutes les traductions, même pour des personnes qui ne connaissent pas le russe. On a comparé une édition de la sortie-machine avec l'original russe et on a vu qu'il n'y avait pas d'erreurs de concept, il y avait des formulations différentes.

Delatte

C'est-à-dire qu'au fond il y a au niveau de l'humain une seconde traduction qui rétablit automatiquement les incorrections de la traduction automatique.

Perschke

Le but principal, c'est la traduction automatique, l'information rapide d'un spécialiste dans son champ d'activité. D'ailleurs, si un linguiste commence à lire un livre sur l'énergie atomique, il ne le comprend pas même s'il est en anglais ou en français parfaits. En revanche, s'il connaît parfaitement la matière traitée, il comprend aussi si elle est en français ou anglais incorrects.

Busa

Quelles sont les spécialisations scientifiques couvertes par ces traductions?

Perschke

La physique,les mathématiques,un peu l'économie. A Georgetown on avait élaboré un corpus de 1.500.000 mots en économie, météorologie, biologie et chimie organique. On a fait une recherche particulièrement développée sur la terminologie des composés chimiques.

Busa

Bien qu'il s'agisse de mots composés chimiques longs?

Perschke

On a la possibilité de détacher avec un tableau des préfixes standard, utilisés aussi pour les composés chimiques.

Duro

Dans votre dictionnaire, est-ce-que la correspondance rassemblée est toujours faite mot par mot, ou plutôt, en plusieurs cas, phrase par phrase, ou locution par locution, groupe binaire, ternaire, quaternaire?

Perschke

Il y a une correspondance mot à mot; quand le mot est polyvalent, il faut faire une analyse syntaxique de la phrase et on pose les conditions, pour le choix des traductions, des valeurs du mot selon la structure syntaxique de

la phrase où il se trouve. On ne tient pas encore compte des informations sémantiques, mais pour la plupart des homographies il est suffisant de trouver les conditions syntaxiques.

Busa

M. Perschke voudrait nous donner des exemplaires de traduction du rus-en anglais scientifique, faite à la machine avec le programme d'Ispra.

Perschke

C'est une traduction d'articles sur la physique du plasma, et je ne crois pas que le sujet de cet article vous intéresse beaucoup, mais c'est une illustration de ce qui sort de l'ordinateur.