World Patent Information 35 (2013) 183–186

Contents lists available at SciVerse ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

The first steps in developing machine translation of patents

L.G. Kravets

"Patent Information Today" Magazine (Russian), Druzinnikovskaya Str. 11A, Moscow 123995, Russia

Keywords:
Machine translation
Patent claims
English text segmentation
Nominal word groups analysis
Russian equivalents synthesis
Soviet Union
Historical

ABSTRACT

The TSNIIPI MT experimental system, developed in 1963–1966, was focused on the translation of publications of the US weekly "Official Gazette, specifically the first paragraphs of patent claims. These claims are characterized by an abundance of difficult to grasp multicomponent terminological combinations and by a specific syntactic structure of unusually long sentences containing up to several hundred words. The system's algorithm performed the segmentation of the English text, and the identification and structural analysis of multicomponent word groups necessary to synthesize the corresponding Russian equivalents.

1. Introduction

The dramatically increased flow of patent documentation coming recently from Asia – especially from Japan, China and Korea – and building a single patent system in the multilingual European Union have concentrated the patent world's attention on the problems of overcoming language barriers with the use of machine translation (MT). Today, this attention is focused basically on the Patent Translate System, which is the result of cooperation between the European Patent Office and Google. Under the agreement, the EPO will use Google's machine translation technology to translate patents into the languages of the 38 countries that it serves. In return, it will provide Google with access to its translated patents, enabling Google to optimize its machine translation technology.

Google Translate is based on a method called *statistical machine translation*, developed by F.J. Och who won the DARPA contest for speed machine translation in 2003 [1]. It takes a statistical approach, comparing the source document sentence by sentence to millions of patent documents previously translated by humans. These are used to train the translation engine to handle technical subject-matter and the specific style and format used for patent documents. The service is certainly useful for getting the gist of a patent written in a foreign language and is helpful for companies attempting to get an informal feel for the competitive patent landscape.

The Patent Translate, used at the EPO, is said to be a machine translation service specifically "trained" to handle elaborate patent vocabulary and grammar. However, as with Google's general translation tool, the results are said to be still far from perfect. Although machines can automate certain tasks very well, none seems yet to have fully mastered the subtle differences in sentence structure and the potential multiple uses of a word to have different meanings in different contexts. Because Google Translate uses statistical matching to translate rather than a dictionary/grammar rules approach, translated text can sometimes include apparently nonsensical and obvious errors, such as swapping common terms for similar but nonequivalent common terms in the other language, as well as inverting sentence meaning [2,3].

By their very nature patents are concerned with new inventions. They will therefore contain new terms, used by inventors to describe their innovations. Consistency of terminology is crucial when creating a patent specification. And there remains a very complicated problem of translating patent claims. They use formalistic language with an unusually long sentence structure, required for clear display of technical and legal aspects of the invention, subject to the broadest possible legal claims. For a machine this is a major problem to overcome [4].

Meanwhile, attempts to solve some of these problems began half a century ago in Moscow, at the Central Research and Development Institute of Patent Information (TCNIIPI), which was entrusted with the processing of foreign patent documents. It was decided to translate into Russian the claims or abstracts published in official bulletins of leading patent offices. Therefore, in parallel with the traditional processing of current patent documentation, the TCNIIPI scientists developed in 1963–1966 an experimental system to automatically translate publications from the USPTO "Official Gazette".

MT development at the TSNIIPI covered a period when – after the thorough theoretical research and the emergence of more efficient computers – several groups around the world had begun their attempts to create practically operational MT systems.

E-mail address: kravets27@yandex.ru.

0172-2190/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wpi.2013.02.002

Confrontation between the two opposing social systems had led to the situation that, by the time MT projects were implemented, they were mostly aimed at providing translation from Russian into English and vice versa. One of them was the first MT system specialized for processing patent texts [5,6].

Subsequent sections of this paper are devoted to the consideration of the linguistic specificity of patent claims, manifested in the predominance of nominative word groups. This places special demands on the MT algorithm, called on to carry out the segmentation of the claims text, the identification and analysis of nominal word groups in the English text and the formation of the equivalent word combinations in the Russian language. The paper ends with a summary of the MT system structure as a whole.

2. Special features of patent claims

Initial attempts to solve the problem facing TSNIIPI by automating the word for word translation of patent claims confirmed the unsuitability of such an approach [7]. Therefore in was decided to develop an MT system with the ability to navigate in the original patent documents [8].

Thorough linguistic analysis of patent claims in the "Official Gazette" showed that the overwhelming majority of the notions and concepts used to describe the basic idea of an invention are expressed by terms which are nominal word combinations with prepositive attributes. The number of such word combinations is practically unlimited, and therefore no automatic vocabulary was able to cover even an essential part of such word groups. This problem becomes still more complicated when translating patent texts in which, due to their specific character (first communication about new inventions), new and derived terms are bound to occur. Careful analysis of nominal word combinations was a prerequisite for improving the quality of translating patent claims [9].

The determining role of different nominal groups in a patent claim influenced the choice of the fundamental principle and construction of a specialized MT algorithm. It was called *the algorithm of segment analysis*. The name reflects the main idea of the algorithm, which provides the division of the claims text on segments, identifies patterns of these segments, finds equivalent models of the Russian language, then develops the information on the grammatical form of Russian words and synthesizes the Russian text in accordance with this information.

The role of segment separators was performed by a number of words: indisputable (e.g., prepositions) and questionable (such as determinatives, unions, participles). If the separator is controversial, analysis of its environment was performed. Thus, the union and the article were not separators if they were between similar definitions.

The text of patent claims in the "Official Gazette", with up to a few hundred words, is designed in the form of a single sentence, which complicates the understanding of the invention. Therefore, an attempt was made to develope formalized rules of dividing continuous text into segments and designing them in separate sentences. Here sentence separators were used too, followed by the analysis of their environment in case of controversy. When presenting separate parts in the form of independent phrases the participles of absolute participle constructions were converted to finite verbal forms. A noun or nominal group being a part of the invention (at itemization) was considered to be subjects, and before them the predicate «imeetsya» ("there is") was inserted.

The analysis of segments was intended primarily to establish the relationships between the words of the English text. If the relationships between the words within the segment are known, it becomes possible to determine the character of relationships between the units of the equivalent Russian segment [9].

3. Identification and analysis of nominal word groups

High quality work with multicomponent noun phrases is largely determined by objective criteria of identifying their structure, otherwise correct clarification of the lexical meaning of complex entities is impossible. Therefore it was decided to use in the MT system the probability analysis of the phrases' structure based on some statistical data quite regularly identifying the types of structural and the corresponding semantic relationships. Admissibility of probability estimates in identifying structural models of multicomponent combinations was tested on a sample of technical texts, which contained about 20,000 two-component and about 5000 multicomponent phrases. Based on this analysis, the diversity of nominal groups reduces to a finite set of models that reflect a summary of their structure and composition. These structural models helped to identify some objective signs that quite regularly point to the relative degree of stability of relationships between the components of the word combination.

The automatic analysis of nominal groups in machine translation was preceded by their identification in the text. Usually, the left boundary of the group was indicated by an article or any other word that acts as a determinative. The right boundary was defined by the core noun itself. The role of prepositive elements of nominal groups may be played by the words in the following classes: defining words – adjectives, participles, pronouns, ordinal numbers (M), nouns (N), adverbs (D) and cardinal numbers (Nu).

In order to automatically analyze the claims all recorded nominal groups were combined into a finite set of *structural models*. Structural model is a category, representing, first of all, two related concepts: a) a *distributive model* – the sequence of the above indices of classes/subclasses of words, which include components of the nominal group, b) a *constructive model* – the type of syntactic connections between components of the group. These were later supplemented by *a semantic model* representing the type of generalized semantic relations between the components of the word group [10].

Two-component word combinations had one of the following three distributive models: MN, NuN and NN. The analysis of three-component word combinations appeared a great deal more complicated because of the increased number of required versions for analysis. Thus on the level of word classes the following 7 distributive models have been established: MMN, DMN, MNN, NMN, NuMN, NuNN and NNN. In case of four-component word combinations the number of distributive models amounted to 15 and so on. Word combinations with the number of components greater than 8 were not analyzed and translated word for word. Since they occurred very rarely (less than 1% of the total number of word combinations) it did not essentially impair the quality of the translation.

The increase in the number of components raises the complexity of a nominative group automated analysis significantly. This was caused by the increasing diversity of syntactic relations. As a result a three-component nominative group may have different constructive models. When analyzing a three-component nominative groups with a distributive model MNN it is necessary, above all, to choose a noun, which is consistent determiner M. For example, the distributive model MNN can match the constructive model ((xy) x) – *internal combustion engine* or (x (yz)) – *additional fuel pump*.

Prior to the operation of the basic blocks of text analysis, grammatical homonymy of words was eliminated by analyzing the grammatical characteristics of the surrounding words. For example, a verb cannot be directly preceded by an article.

4. The synthesis of the Russian text

In accordance with the adopted structure of the algorithm, the basic information required to obtain correct grammatical forms of