Automatic identification of the semantically equivalent fragments of the text documents

Yury Krapivin, *Brest State Technical University*
(**01.11.2008,** prof. Vladimir Golovko, *Brest State Technical University*)

Abstract

The article presents a definition of plagiarism identification problem and its solution via automatic identification of the semantically equivalent fragments of the text documents using SAO relations retrieved from the texts.

1. Problem definition.

The problem designated as the name of the article is considered in the context of one important application – automatic plagiarism identification, that usually refers to intentional appropriation of another authorship of literary, scientific, art, invention work or innovation proposal (fully or partly). Plagiarism events can also be unpremeditated, for example, due to strong external informational influence that can refer to using ideas or distinctive expression style as well as disregarding quote standards, in case of textual expression of information. Thus, it is reasonable to consider the implementation of mentioned application as a sequence of two following steps:

- the identification of equivalent, in certain sense, fragments of the given text document and text documents from the given database and available from the Internet–resources;
- the analysis of equivalent fragments in terms of their adoptions, with the involvement of experts, i.e. regarding plagiarism identification.

Upon the analysis of the problem, in the case of equivalent textual fragments, in this sense, certainly, focus should be made on completely coinciding fragments as well as those that match up to some criteria determined by premeditated and quite easy actions (procedures), taken by the authors of texts on purpose of transferring the problem of plagiarism identification from the purport using relatively simple indicators to compare text fragments to another one using indicators derived via serious linguistic analysis, i.e. with a view to complicate the solving the problem. Those procedures may include the following:

- permutations of words permitted in terms of the language grammar;
- (not) using uninformative words, e.g. parenthetical constructions;
- using synonyms of words for particular parts of speech (nouns, verbs, articles and so on), voice synonyms and different synonymous constructions on noun-phrases level, object-parametrical relations (for example “heat A” = “step up the temperature of A”) etc.;
- using paraphrase, i.e. text fragment narration that retains its basic meaning.

It should be noted, that the later of the listed procedures is based, inter alia, on the set of antecedents. In regard to the main meaning of the text fragment, it can be treated, for example, as the set of these knowledge that exist there, referred to three main classical knowledge types [1]: objects/object classes, facts (semantic relations as S-A-O, where S refers to a subject, A – to an action, O – to an object) and rules (cause-and-effect relations between facts per se), that represent regularities of environment/knowledge domain.

In consideration of cause-and-effect relations that operate with facts and facts – with objects for the considered problem obviously, it’s possible to confine ourselves to the second type of knowledge. Thus, we will consider two text fragments semantically equivalent, if their sets of facts match up to the synonymy of the units they consist of. And it thus comes to the automatic identification of these fragments in text documents.

2. Resolving the problem.

The relevance of identification of the completely coinciding text fragments was noted in the previous section of the article. The problem was resolved in this statement for texts in Russian and Belorussian [2], but algorithms we have designed are suitable for different languages. In regard to the identification of the semantically equivalent text fragments, i.e. resolve the problem in abstract definition, it, obviously, will require the availability of the linguistic processor (LP) referred to automatic analysis of the text in all depth levels of language – from lexical up to semantic one. A well-known multilingual LP [1] was used as such one. Text virtually in any currently used formats (DOC, PDF, RTF, HTML, XML, TXT and so on) goes to its input and is stepwise processed via preformatting, lexical (recognition of word and sentences boundaries), lexico-grammatical, syntactical and semantic analyse. At the last step, particularly, so called extended facts are identified, i.e. semantic relations as SAO (Fig. 1, is illustrated by the example of English).

Component name	Definition
Subject	*subject*, concept performs the action (water is heated by **fire**)
Action	*action*, performed by the subject to the object (the workers **build** a house)
Object	*object*, concept recipient of the action (**house** is built by the company)
Adjective	attribute of the action – *adjective* (the invention is **efficient**; the water becomes **hot**)
Preposition	circumstance of the action or object – *preposition*, usually coupled with *indirect object* (the lamp is placed **on** the table)
Indirect Object	*indirect object* of the action, often coupled with *preposition* (the lamp is placed on **the table**)
Adverbial	attribute of the action with *adverb* function (the object is **slowly** modified; the driver must not turn the steering wheel **in such a manner**)

Fig.1. A structure of the semantic relation SAO

It's quite natural that certain components of the relation during the identification of SAO in concrete sentences of the text document can be void, for example SAO components Subject, Adjective and Adverbial from the sentence "the lamp is placed on the table" are void by reason of the given sentence structure (Fig.2).

Component name	Definition
Subject	-
Action	place
Object	lamp
Adjective	-
Preposition	on
Indirect Object	table
Adverbial	-

Fig.2. An example of the partially filled semantic relation SAO from the sentence "The lamp is placed on the table."

It's obvious, that different syntactic structures, that express equal or similar meaning, can correspond to "fact"-type knowledge in text. Thus, the fact "fire-heat-water", that is identified in the phrase "fire heats water", can also be represented with other syntactic forms:

- water is heated by fire;
- fire is able to heat water;
- using of fire allows to heat water;
- heating of water is accomplished with help of fire.

We, obviously, ensure the resolving of the considering problem, via the supplement of the linguistic database of the mentioned LP with dictionaries of parenthetical constructions and synonyms for the particular parts of speech, defined as component structure of the extended fact, and its functionality – with appropriate retrieval procedures trough this dictionaries.

In regard to the algorithm of the identification of the semantically equivalent text fragments its schematic diagram is, actually, analogous to the algorithm of the recognition of the adopted sentences represented in [2], on the assumption of considering the text documents as the sequence of facts, not words. The conditions of both complete and partial matching of those sequences by the equal facts percent of their total amount in the chain, as well as by the component structure of the comparing facts, and filling equal components can be specified.

One of the results of identification of two semantically equivalent text fragments produced by the prototype system is shown bellow.

Fragment 1.

... A laser is a device that emits light through a process of optical amplification based on the stimulated emission of photons. A laser consists of a gain medium and optical cavity for providing the optical feedback. The light that is emitted by the laser is notable for its high degree of spatial and temporal coherence...

Fragment 2.

... A device that is able to emit light by means of a process of visual amplification that is based on the photons emission is called laser. A gain medium and optical cavity to provide optical feedback are main parts of laser. The light emitted by the laser is known for high degree of temperature and spatial coherence...

Shown text fragments consist of semantically equivalent sentences. Thus, for example, after processing via LP first sentences of the shown fragments, appropriately next facts will be identified there:

$F_1^{(1)}$ laser – be – device

$F_2^{(1)}$ laser – emit – light – **through** – process of **optical** amplification

$F_3^{(1)}$ X – base – process of **optical** amplification – on – **stimulated emission of photons**

$F_1^{(2)}$ laser – be – device

$F_2^{(2)}$ laser – emit – light – **by means of** – process of **visual** amplification

$F_3^{(2)}$ X – base – process of **visual** amplification – on – **photons emission**

Synonymous components of corresponding facts are marked out here: "through" from $F_2^{(1)}$ and "by means of" from $F_2^{(2)}$ and so on. Indirect objects "stimulated emission of photons" ($F_3^{(1)}$) and "photons emission" ($F_3^{(2)}$) are recognized as synonymous (conditionally) owing to noun phrase synonymy criteria accepted in this version of the system (with no account taken of attribute is permitted). Fixation "laser" as the subject of the facts $F_2^{(1)}$, $F_1^{(2)}$ and $F_2^{(2)}$ is possible owing to the presence in using LP anaphora resolution functionality. The "void" subject in mentioned facts is marked with the sign "X".

3. Conclusion.

The results submitted above were successfully implemented as a prototype system that is used for recognition in text documents semantically equivalent text fragments. The submitted solutions allow the system built according to them to recognize the explicit and implicit adoptions referring to the knowledge of a natural language up to the semantic level.

4. Bibliography And Authors

[1] Совпель И.В.: *Система автоматического извлечения знаний из текста и её приложения* // Искусственный интеллект. – 2004. – Т.3. – С. 668-677.

[2] Крапивин Ю.Б.: *Автоматический поиск заимствованных из Интернет-источников фрагментов* // Искусственный интеллект. – 2012. – Т.4. – в печати.

Authors:

Mr. Yury Krapivin
Brest State Technical University
Moskovskaja str., 267
224017 Brest, Belarus
tel. +375 297 98 81 46
fax +375 162 42 21 27
email: *ybox@list.ru*