Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 23 (2009) 107–125

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Natural language watermarking via morphosyntactic alterations

Hasan Mesut Meral [a], Bülent Sankur [b], A. Sumru Özsoy [a,c], Tunga Güngör [c,d,*], Emre Sevinç [c]

[a] Boğaziçi University, Linguistics Program, Bebek, İstanbul 34342, Turkey
[b] Boğaziçi University, Department of Electrical and Electronic Engineering, Bebek, İstanbul 34342, Turkey
[c] Boğaziçi University, Cognitive Science Program, Bebek, İstanbul 34342, Turkey
[d] Boğaziçi University, Department of Computer Engineering, Bebek, İstanbul 34342, Turkey

Received 17 October 2007; received in revised form 17 April 2008; accepted 28 April 2008
Available online 3 May 2008

Abstract

We develop a morphosyntax-based natural language watermarking scheme. In this scheme, a text is first transformed into a syntactic tree diagram where the hierarchies and the functional dependencies are made explicit. The watermarking software then operates on the sentences in syntax tree format and executes binary changes under control of Wordnet and Dictionary to avoid semantic drops. A certain level of security is provided via key-controlled randomization of morphosyntactic tools and the insertion of void watermark. The security aspects and payload aspects are evaluated statistically while the imperceptibility is measured using edit-hit counts based on human judgments. It is observed that agglutinative languages are somewhat more amenable to morphosyntax-based natural language watermarking and the free word order property of a language, like Turkish, is an extra bonus.

Keywords: Natural language watermarking; Tree bank; Agglutinative; Morphosyntax; Text payload

1. Introduction

Natural language watermarking (NLW) is an emerging research area at the intersection of natural language processing and information security. It aims to hide information in texts with applications similar to those in multimedia watermarking (Cox et al., 2002). The goals could be to create a subliminal communication channel through which to transport hidden information, to enable content and authorship authentication, to enrich the text with metadata, to fingerprint it for distribution, etc. While natural language watermarking and multimedia watermarking share common goals, they employ very different techniques. A plethora of watermarking techniques have been explored for multimedia documents in the last decade (Cox et al., 2002) and some have even turned into industrial products. In contrast, studies on natural language watermarking are just starting as

[*] Corresponding author. Tel.: +90 212 3597094; fax: +90 212 2872461.
E-mail addresses: mesut.meral@boun.edu.tr (H.M. Meral), bulent.sankur@boun.edu.tr (B. Sankur), ozsoys@boun.edu.tr (A. Sumru Özsoy), gungort@boun.edu.tr (T. Güngör), emres@bilgi.edu.tr (E. Sevinç).

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.04.001

attested by the scarcity of related papers (Bergmair, 2004, 2007; Khankhalli and Hau, 2002; Bennett, 2004).

Initially, NLW researchers exploited watermarking techniques adapted from multimedia watermarking. These watermarking and/or steganographic techniques were non-linguistic in nature and made extensive use of character changes such as kerning, random assignment of character spaces, line shifting, word shifting and insertion of sound encoding (Bailer and Rathner, 2001). These "printed text" watermarking approaches had limited scope and were not robust against text reformatting and transcription attacks (Khankhalli and Hau, 2002). Since the exploitable redundancy and embedding opportunities in printed text are significantly less than in multimedia documents such as image, video, audio and graphics, attention was turned to linguistic tools. These tools can appear under the guise of semantic and syntactic transformations, morphological and punctuation manipulations, lexical substitutions, translations and word level typographical alterations (Topkara et al., 2005; Meral et al., 2006). We will refer to the term "natural language watermarking" as the information hiding techniques within a text exclusively based on linguistic tools, while the term "text-watermarking" loosely encompasses both document image formatting and linguistic manipulations.

The motivation for our work is to develop a novel NLW scheme that is imperceptible, secure and based on morphosyntactic manipulations of sentences. We develop our scheme based on the seminal work of Topkara et al. (2006b). The contributions of our work can be summarized as follows: (i) we analyze a fairly complete list of morphosyntactic tools. We observe that agglutinative languages with high suffixation, such as Turkish, constitute fertile ground for watermarking; (ii) our scheme allows trade offs between payload and security in a controlled manner, and these properties are quantified based on a language model and its statistics; (iii) imperceptibility or acceptability of manipulations is measured via edit statistics.

The morphosyntactic approach was chosen de facto since alternative approaches did not look very viable. Most languages in contrast to English are not very rich in synonyms; in fact, some languages like Turkish have a single word per concept, and borrowed synonyms from other languages look alien. Purely morphological watermarking was not considered since morphology and syntax work often hand in hand, hence should be better handled under the guise of morphosyntax. Watermarking via punctuation alterations is not a stable enough option and can easily result in unwanted stylistic and meaning differences. Semantic watermarking, as argued by Atallah et al. (2000), constitutes probably the most flexible and prolific approach to NLW. However, the present day technology does not yet offer adequate tools for semantic interventions and pragmatic extensions in the word domain leading to semantic watermarking. On the other hand, first, the morphosyntax of languages often offers a rich set of NLW tools (Topkara et al., 2006b). Second, the syntactic alterations are based on formal descriptions of linguistic expressions in a sentence domain for which parsers and transformers exist or can be built. Finally, we conjecture also that the manipulation of the morphosyntactic features would have less impact on the semantics (i.e. least semantic distortion) of the original text when compared with the alternatives of lexical and semantic feature transformations.

Although syntax-based NLW has been proposed before, the algorithm proposed here is novel in its pseudo-random recruitment of morphosyntactic tools and subjective assessment of watermarked texts. The proposed NLW algorithm is applicable to any language, given a repertoire of morphosyntactic alternative forms, and ancillary material such as Wordnet and Dictionary. It processes a text progressively at the sentence level and selectively implements feasible watermarking tools when it encounters their appropriate input representation. As a case study, imperceptibility and payload capacity are given for Turkish language, which is an agglutinative language rich in morphosyntax. In order to explore the watermarking potential of this agglutinative language, we introduce a two-level embedding algorithm, which takes as input a dependency tree and converts it into a structure representing hierarchical relations. To enhance robustness of the watermark we introduce randomized order of tool selection and insertion of a "pass" tool creating void watermarks. Finally, we measure imperceptibility via user feedback.

This paper is organized as follows: In Section 2, we discuss the current state of the art in NLW and we expound the methodological basis of our study. In Section 3, we introduce the NLW model and describe the functions of its modules. Section 4 provides the results of the watermarking experiments where we discuss the occurrence statistics, imperceptibility effects, security measures and payload. Conclusions are drawn in Section 5. Appendix A gives the NLW tool repertoire for Turkish and English.