Два метода синонимического перефразирования в лингвистической стеганографии

И. А. Большаков
Центр Компьютерных Исследований
Национальный Политехнический Институт (IPN)
Мексика, 07738, г. Мехико
igor@cic.ipn.mx

Стеганография – это сокрытие одной информации в другой. Сам факт внедрения шифровки скрывается, а несущая информация должна сохранять безобидность и осмысленность. Лингвистическая стеганография (ЛС) скрывает один текст в другом, опираясь на свойства языка и лингвистичекие ресурсы. Для целей ЛС мы предлагаем два метода синонимического перефразирования текста-носителя. Метод синонимическиз замен заменяет отдельные слова текста их синонимами. Абсолютные синонимы используются независимо от контекста, а относительные проверяются на совместимость заменяющего слова с контекстными словосочетаниями. Конкретная замена определяется шифруемой информацией. Словосочетания – это синтаксически связанные и семантически совместимые пары полнозначных слов. Они собираются заранее в обширную базу словосочетаний, с которой взаимодействует словарь синонимов. Оба эти ресурса содержатся в системе КроссЛексика. Метод синонимических перестановок упрощенно анализирует предложения несущего текста, выявляя в них составляющие верхнего синтаксического уровня. Из заранее составленного словаря допустимых перестановок составляющих берется та, что соответствует шифруемой информации. Метод замен описан детально, метод перестановок лишь намечен. Обеспечиваемая стенографическая плотность для обоих методов равна примерно 1/200. Методы независимы и совместны, а при совмещении их показатели суммируются.

Введение

Стеганография тайно передает одну информацию в другой. При этом сам факт передачи секрета не скрывается. Поскольку объем чисто текстовой информации в каналах связи и интернете постоянно растет (текущие новости, электронная почта, реклама, спам и пр.), есть надежда передавать секретные сообщения прямым их внедрением в несущие тексты, пусть с невысокой скоростью, но незаметно. Этим и занимается текстовая стеганография. Ей важно сохранить текст-носитель безобидным и осмысленным.
Безобидным считается текст, не привлекающий внимание внешней стороной – форматом, шрифтом, орфографией, морфологией, синтаксисом или лексикой. Все эти черты должны соотноситься лишь с темой текста. Осмысленный текст последовательно излагает нечто отличное от спрятанного сообщения. Если смысловая связность носителя при кодировании теряется, то утрачивается и его безобидность. В идеале носитель сохраняет исходный смысл.

Для внутритекстового кодировании первой приходит идея использовать формат, но тогда переформатирование полностью стирает все скрытое. Более практична собственно лингвистическая стеганография (ЛС) – внедрение произвольной информации в произвольный текст с опорой на свойства языка и лингвистические ресурсы.
В [6] для ЛС используются словари квазисинонимов. Лексика разбивается на множество групп разного объема. Внутри групп слова сходны как грамматически, принадлежа одной части речи, так и семантически, вплоть до настоящей синонимии. Если очередное слово несущего текста принадлежит группе с $m > 1$ синонимами, оно может нести скрытую информацию. Пусть размеры всех полученных групп кратны степени двойки или сокращены до ближайшей степени двойки. Тогда для каждой группы размера 2^n из кодируемого сообщения выделяется слог длины n и его двоичное содержимое берется в качестве внутригруппового номера синонима, подставляемого в текст вместо исходного. Операция на приемном конце очевидна. Но при крупных группах ничем не ограниченные замены делают носитель невразумительным.
В предлагаемых нами методах тоже используется синонимическое перифразирование, но разного типа. Феномен синонимии очень важен в лингвистике. Так, теория «Смысл ⇔ Текст» считает себя исчислением синонимических перифразов [8]. В ней предложена сложная совокупность преобразоваиий, сохраняющих смысл предложения и дискурса в целом. В зависимости от того, на каком уровне языка ведется перефразирование, в его процессе меняется лексика, синтаксическая структура, морфологические характеристики слов, их число и порядок. Однако программное воплощение теории пока затронуло немногое в языке [1].
Согласно нашему методу замен, как и в [6], производятся синонимические замены, сохраняющие порядок слов, синтаксическую структуру предложения, и, приблизительно, число слов в нем. Производимые замены тестируются относительно контекста: проверяется, входит ли заменяющее слово в набор словосочетаний заменяемого слова. Только если данная замена контекстно допустима, соответствующий синоним оставляется в группе потенциальных замен. Конкретная замена определяется кодируемой информацией.
Словосочетания – это синтаксически связанные и семантически совместимые пары полнозначных слов,| например, *правильно выразить, передать по радио, глава государства.* Предполагается, что измеряемые сотнями тысяч словосочетания произвольной частотности и идиоматичности собраны заранее в некую базу, где синонимы ищутся как компоненты словосочетаний.
Согласно нашему методу перестановок делаяются синонимические перестановки составляющих предложения. Для выделения составляющих верхнего синтаксического уровня (напр., обстоятельств времени и места) использует упрощенный синтаксический анализ. Далее по заранее составленному словарю допустимых перестановок выявляется, такие перестановки допустимы для данного набора составляющих и берется по номеру соответствущая кодируемой информации.
Методы могут применяться независимо и совместно, сохраняют исходный смысл текста, а обеспечиваемые ими показатели плотности кодирования при совмещении складываются.

Абсолютные и относительные синонимы

Синонимы – это слова, могущие замещать друг друга в некотором классе контекстов с незначительным изменением смысла полного текста. Обороты с «некоторый» и «незначительный» делают данное определение нечетким, но синонимические словари продолжают строиться на его основе. Типичный синонимический словарь состоит из групп слов, считаюшихся синонимами друг другу. Обычно выделяют титульное слово группы

(доминанту), выражающее смысл группы наиболее общим и нейтральным способом. Каждое слово синонимической группы может иметь смысловое сходство и с иной группой, даже быть в нее включенным, т.е. группы могут пересекаться.
Мы не ограничиваемся лишь однословными синонимами, нам подходят даже группы без единого однословного члена: {*надеяться, возлагать надежды, питать надежду*}, {*наконец, в конечном счете, в конце концов*}, {*в течение нескольких дней, за несколько дней* } и т.п.
Очень важна абсолютная синонимия. Она не меняет смысла текста при любых контестах (*лингвистика = языкознание*). К сожалению, абсолютных синонимов редки, но много эквивалентов иного рода – различных сокращений. Вот группа эквивалентов: {*Соединенные Штаты Америки, Соединенные Штаты, США*}. Многословные синонимы привносят их множество: {*экс-президент, бывший президент*}, {*замминистра, заместитель министра*}. В языке интернетовских новостей используется несколько тысяч склеек типа *детсад = детский сад*, *сейсмостанция = сейсмическая станция*, *физлица = физические лица*. Есть еще так называемые морфологические варианты типа {*нуль, ноль*}. Будем относить всех их к абсолютным синонимам.
Неабсолютные синонимы назывем относительными. Как средство отличения абсолютных синонимов, один из них берем доминантой, а прочие снабжаем пометой.
Итак, предполагается обширный синонимический словарь, где

- Каждая группа имеет доминанту.
- Эквиваленты доминанты, если они есть, помечены.
- Любой член группы может состоять из несколько слов.
- Любой член группы может повторяться в иной группе и/или быть омонимом члена другой группы.
- Члены группы могут характеризовать не полную лексему, а ее граммему, т.е. отдельно ед. и множ. число существительного, отдельно личные формы глагола + инфинитив, причастия и деепричастия, да еще и разделенные на совершенный и несовершенный вид. (О необходимостим граммем см. ниже.)

База словосочетаний и дополнения к ней

Словосочетания соединяют слова синтагматически, например, глагол и валентное ему существительное или существительное и определяющее его прилагательное. В русском языке словосочетания в всегда брались в данном понимании. В рамках теории «Смысл ⇔ Текст» они получили адекватное описание и классификацию [9]. Семантически они делятся на фраземы (их полный смысл не включает прямой смысл компонентов); полуфраземы (содержат прямой смысл только одного компонента) и свободные сочетания (смысл составлен только из смыслов компонентов).
В 1990-2003 годах нами была разработана интерактивная система КроссЛексика, в базе которой содержится ныне более 1,2 млн русских словосочетаний разной частотности и идиоматичности – фразем, полуфразем и свободных словосочетаний [2, 4, 5]. Статистика показала, что свободные словосочетания не столь уж свободны: они возможны только между словами фиксированных семантических полей, и в целом их больше, чем фразем и полуфразем вместе взятых, лишь раз в пять. Именно свободные словосочетания обеспечивают подобным базам широкое применение для создания гибкого и идиоматичного текста, совершенствования синтаксического анализа, разрешении омонимии, обнаружения и исправления семантических ошибок, синонимического перефразирования и пр.
Примерно 94% словосочетаний в КроссЛексике оказались следующими: существительное / глагол / прилагательное / наречие – его модификатор (прилагательное или наречияе) (*запутанный сюжет*, *правильно выразить, практически незаметный*); глагол – его подлежащее (*сушествует противоречие*); глагол – его дополнение или предложное обстоятельство (*дать воды*, *передать по радио*); существительное – его дополнение (*глава*

государства). В качестве компонентов словосочетаний берутся не лексемы целиком, а их морфологические подпарадигмы, называемые граммемами. Это обстоятельство требует пояснения.

Было обнаружено [3], что существительное в ед. и множ. числе может иметь разные наборы словосочетаний. Поэтому граммемы двух чисел взяты разными единицами словаря. Глаголы играют разные синтаксические роли: сказуемого (в личных формах), определения (в форме причастия), обстоятельства (в форме деепричастия) и имеют словосочетания разных типов. Поэтому рассматриваются раздельно все личные формы + инфинитив, причастие (походит на прилагательное) и деепричастие (походит на наречие). Дополнительное членение производится и по признаку вида глагола.

Далее полагается, что алгоритм синонимических замен пользуется при тестировании полной базой словосочетаний КроссЛексики. Кроме словосочетаний здесь содержатся и семантические связи ВордНетовского типа [7], из которых привлекаются синонимы и гиперонимы. Синонимический словарь оперирует теми же граммемами, что и база словосочетаний.

Множество словосочетаний вне базы КроссЛексики можно получить эвристическим «выводом». Так, в базе имеются словосочетания о цветах вообще {*купить цветов, украсить цветами,...*}, и известно, что калы являются подвидом цветов. Поэтому можно «вывести» словосочетания {*купить кал, украсить калами,...*} [4].

Алгоритм синонимических замен

Входами с предлагаемый алгоритм служат двоичная информация, предназначенная для шифровки, и несущий текст на русском языке, по объему примерно в 200 раз больший, чем у шифруемой информации. Формат текста произволен, но он орфографически и синтаксически правилен, дабы не спровоцировать исправлений при передаче. Последовательности цифр или личных имен допускаются, но они увеличивают требуемую длину текста. Алгоритм включает следущие шаги.

Поиск синонимичных слов. В тексте отыскиваются слова и многословные выражения, имеющие синонимы. Если одновременно найдена последовательность слов и ее подпоследовательность, предпочтение отдается объемлющей.

Формирование объединенных синонимических групп. Последовательно рассматриваются синонимичные слова текста. Если в соответствующей синонимической группе только абсолютные синонимы, она принимается безоговорочно. Если в группе есть хоть один относительный синоним, все такие синонимы подвергаются операции транзитивного замыкания. При замыкании для каждого синонима проверяется, не является ли он членом какой-либо иной синонимической группы. Если это так, дополнительная группа присоединяется к исходной без повторов. Далее присоединенные синонимы просматриваются на принадлежность к иным, еще не рассмотренным синонимическим группам, и так до исчерпания пополнений. Замыкание совершается также через омонимы. Анализируется, не является ли омонимичным исходное текстовое слово или какой-нибудь член его синонимической группы. Если это так, и если еще не рассмотренный омоним имеет синонимы, привлекается группа синомнимов этого омонима. Каждая вновь привлеченная группа используется для расширения и т.д. Процесс конечен, но иногда дает обширную объединенную группу. Транзитивное замыкание необходимо, поскольку делает состав объединенной группы не зависящим от того, с какого члена замыкание начинается.

Проверка словосочетаний. Если группа содержит только абсолютные синонимы, она не проверяется на контекст, а для проверки на сочетаемость с ней иных слов может браться любой ее член. Если же группа имеет относительные синонимы s_i, они подлежат проверке на совместимость с внешними полнозначными словами w_j слева и справа от проверяемой

группы. Если внешнее слово w_j не синонимично или имеет только абсолютные синонимы, то проверяется, с какими из s_i оно образует однотипные словосочетания. Те s_i, которые не формируют словосочетаний с w_j, отбрасываются. Если внешнее слово w_j само принадлежит объединенной группе с элементами w_{jk}, то проверяются все однотипные словосочетания пар $\{w_{jk}, s_i\}$ при всех i и k. Отсутствие словосочетания хотя бы с одним внешним элементом ведет к отбрасыванию проверяемого элемента. При этом разных пар может не остаться и тогда группа в стеганографии не участвует. Элементы, оставшиеся в группе, нумеруются фиксированным образом от 0 до $m - 1$, где m – число оставшихся элементов.

Кодирование. Последовательность профильтрованных групп сканируется. Если их размеры кратны степени двойки или сокращены до ближайшей степени двойки, для очередной группы длиной 2^n из кодируемого сообщения выделяется слог длины n и его двоичное содержимое берется в качестве внутригруппового номера синонима, подставляемого в текст вместо исходного. Та же операция повторяется для всех групп вдоль текста. Если имеются группы, по длине не равные степени двойки, все длины групп перемножаются и берется степень 2^N, ближайшая к полученному произведению вниз. Затем от кодируемой информации отсекается слог длины N и из него путем последовательных делений и нахождений остатков находятся номера омонимов для замены синонимичных слов в тексте. Если текстовой синоним при кодировании оказался замененным, то в общем случае пересогласуются морфосинтаксические характеристики заменителя и контекста.

Пример, протрассированный вручную

Возьмем типичный текстовой фрагмент из потока новостей *Газета.Ру*:

1. *Пять **подземных толчков** зарегистрировано **за сутки** на юге Республики Алтай. Сила **землетрясений** составляла от 2,2 до 3,1 балла по шкале Рихтера, сообщили на Акташской **сейсмической станции** сегодня после полудня.*

Здесь синонимичные слова или цепочки слов, подчеркнуты, а абсолютные синонимы выделены еще и шрифтом.

Вот группы абсолютных синонимов (они упорядочены по алфавиту и двоично пронумерованы):

0. землетрясения *1. подземные толчки*
0. за 24 часа *1. за сутки*
0. сейсмическая станция *1. сейсмостанция*

При транзитивном замыкании синонима *зарегистрированный* выявляются синонимические группы с доминантами *зарегистрированный, закрепленный, помеченный, отпразднованный* и объединенная группа равна

2. *закрепленный, замеченный, зафиксированный, зарегистрированный, отмеченный, отпразднованный, подмеченный, помеченный, прикрепленный, примеченный*

А вот группы относительных синонимов в тексте, транзитивным замыканием не изменяемые:

0. Алтай *1. Республика Алтай*
0. равняться *1. составлять*
0. проинформировать *1. сообщить*
0. во вторую половину дня *1. после полудня*

Существительное *сила* имеет два омонима каждый со своим набором синонимов:

00. магнитуда *01. мощность* *10. мощь* *11. сила$_1$*
0. действенность *1. сила$_2$*

Отфильтруем теперь относительные синонимы, не отвечающие контексту. Абсолютный синоним *землетрясения* образует словосочетания с *замеченный, зарегистрированный, зафиксированный* и *отмеченный*, остальные члены группы (2) отбрасываются. Все

оставшиеся члены сочетаются с абсолютным синонимом *за сутки* и с несинонимичным словом *юг*. В итоге (2) принимает вид

00. *замеченный* 01. *зарегистрированный* 10. *зафиксированный* 11. *отмеченный*

Из групп *сила* только синонимы *сила*$_1$ удовлетворяют контексту. У *сейсмическая станция*, эквивалентного *сейсмостанция*, нет словосочетаний в базе, но их имеет родовое понятие *станция*, так что выводим: (*сейсмостанция* IS_A *станция*) & (*сообщить на станцию*) → (*сообщить на сейсмостанцию*). Группы *сообщить* / *проинформировать* и *после полудня* / *во вторую половину дня* полностью сочетаемы и поэтому сохраняют свой состав.

В целом синонимы в (1) позволяют закодировать 12 бит информации, например, две латинские буквы с кодами в виде правых 6-битовых слогов таблицы ASCII. Так, биграмма *по* соответствует безупречному варианту (вносимые отличия выделены):

Пять подземных толчков зарегистрировано за сутки на юге Республики Алтай. ***Мощность*** *землетрясений составляла от 2,2 до 3,1 балла по шкале Рихтера, сообщили на Акташской* ***сейсмостанции*** *сегодня после полудня.*

Поскольку объем скрытой информации 1,5 байт, а текста – 206 байт, первый составляет 1/135 последнего. Это стеганографическая плотность. Она невелика, и в примерах мы редко находили большую. Но в многокилобайтном тексте можно скрыть нечто вполне содержательное.

Метод синонимических перестановок

Возьмем заголовок из новостей *Газета.Ру*:

1. [*В Иране*]$_L$ [*во вторник*]$_T$ [*произошло*]$_V$ [*новое землетрясение*]$_S$

Набор составляющих верхнего синтаксического уровня в (3) типичен: обстоятельство места L, обстоятельство времени T, сказуемое V и подлежащее S. Если перебрать все 4! = 24 перестановки, то еще три, TLVS, SVTL, TVSL, оказываются вполне эквивалентными (3); 10 цепочек передают тот же смысл, но с иным намерением высказывания (напр., *Новое землетрясение в Иране произошло во вторник*); еще 10 цепочек недопустимы – вообще или в жанре новостей (напр., *Произошло во вторник в Иране новое землетрясение*). Объединим четыре безупречных варианта в группу синонимических перестановок, предположительно верных для всех предложений состава {S, V, L, T}, и поместим их в специальный словарь. При стеганографии предлагается анализировать каждое предложение на подобные составляющие и при выявлении указанного состава отсекать от шифруемой информации двухбитовый слог и брать его содержимое в качестве внутригруппового номера перестановки, осуществляемой в тесте-носителе. Поскольку текст (3) содержит 48 байт, в примере достигается тогда стеганографическая плотность 1/192, но обычно мы получали меньше.

Теперь дадим пояснения и обобщения. Иное намерение высказываний именуется в теоретической лингвистике иным тема-рематическим членением предложения [10]. Пока мы не готовы истолковывать дескриптивные результаты [10] в прикладном смысле и действуем эмпирически. Для более полного учета перестановок в произвольных предложениях новостей необходимо учитывать также два дополнения и обстоятельство цели/причины (*из-за мороза, в результате оползня* и т.п.). Если потребовать, чтобы данная составляющая в предложении не повторялась и никогда не распадалась на две других, нужно еще делить составные сказуемые (*намерен* | *предъявить*), а обстоятельства места разбивать на географические (*в Иране*), локальные (*в аквапарке*) и смешанные (*в московской больнице*). В итоге достаточно примерно десяти классов составляющих и менее сотни их сочетаний в качестве групп словаря перестановок.

К выбору допустимых перестановок можно подойти гибче. Оценим в 5 баллов цепочки, появляющиеся в тексте, в 4 балла – полностью им эквивалентные и в 3 балла – отличные по намерению. Наберем соотвествующую статистику баллов по группам одинакового состава (но не порядка!), напр., в виде LTVS – 4,57; TLVS – 4,14; SVTL – 4,00; SVLT – 3,28. Если установить порог допустимости для всех групп в 3,2, то в данную группу войдет уже пять членов и в ней можно будет закодировать больше информации. Балансирование идет между увеличенной стеганографической плотностью и возможным появлением неверных перестановок. Подробное исследование проблемы только началось.

Выводы

Предложены два метода лингвистической стеганографии, базирующиеся на синонимическом перифразировании и сохраняющие смысл несущего текста, а потому и его безобидность. Обеспечиваемая стеганографическая плотность в каждом методе близка к 1/200. Поскольку методы базируются на почти независимых феноменах, они могут применяться одновременно, и тогда эти показатели суммируются.

Метод синонимических замен требует крупных лингвистических ресурсов – обширной базы словосочетаний и специально подготовленного синонимического словаря. Именно поэтому в данный момент он может быть применен лишь к русскому языку, для которого соответствующие лингвистические ресурсы уже существуют (система КроссЛексика).

Для метода синонимических перестановок требуется упрощенный синтаксический анализатор и особый словарь синонимических перестановок, где членами групп являются цепочки составляющих, грамматически допустимые и сохраняющие намерение сообщения.

Литература

1. Apresian, Ju. D., et al. ETAP-3 Linguistic Processor: a Full-Fledged NPL Implementation of the Meaning–Text Theory. // Proc. First Intern. Conf. Meaning–Text Theory MTT 2003, Paris, Ecole Normale Supériore, 2003.– p. 279–288.
2. Bolshakov, I.A. Getting One's First Million... Collocations. // A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 5th Intern. Conf. CICLing-2004, Seoul, Korea. LNCS 2745, Springer, 2004.– p. 229–242.
3. Большаков И.А., А.Ф. Гельбух. Раздельное представление словосочетаний для существительных единствен-ного и множественного числа. // Труды Международного Семинара по вычислительной лингвистике и ее приложениям. Диалог'96. Пущино, 1996.– с. 42–44.
4. Bolshakov, I. A., A. Gelbukh. Heuristics-Based Replenishment of Collocation Databases. // E. Ranchhold, N. J. Mamede (Eds.) Advances in Natural Language Processing. Proc. Intern. Conf. PorTAL 2002, Faro, Portugal. LNAI 2389, Springer, 2002.– p. 25–32.
5. Bolshakov, I. A., A. Gelbukh. Word Combinations as an Important Part of Modern Electronic Dictionaries. SEPLN Revista, No. 29, 2002. – p. 47–54.
6. Chapman, M., G.I. Davida, M. Rennhard. A Practical and Effective Approach to Large-Scale Automated Linguistic Steganography. // G. I. Davida, Y. Frankel (Eds.) Information security. Proc. of Intern. Conf. Information and Communication Security ICS 2001, LNCS 2200, Springer, 2001.– p. 156–165.
7. Fellbaum, Ch. (Ed.) WordNet: An Electronic Lexical Database. MIT Press.– 1998.
8. Мельчук, И. А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». Семантика, синтаксис. М.: Наука. – 1974. – 314 с.
9. Mel'čuk, I. Phrasemes in Language and Phraseology in Linguistics. // M. Everaert et al. (Eds.) Idioms: Structural and Psychological Perspectives. Hillsdale, NJ/ Hove, UK: Lawrence Erlbaum Associates Publ.– 1995.– p. 169–252.

10. Mel’čuk, I. Communicative Organization in Natural Language. John Benjamin Publ.– 2003.– 393 p.