СИСТЕМА ДЛЯ ЛИНГВИСТИЧЕСКОЙ ОЦЕНКИ ПСИХОЛОГИЧЕСКИХ ПРОФИЛЕЙ[1]

SYSTEM FOR LINGUISTICALLY-BASED EVALUATION OF PSYCOLOGICAL PROFILES

Г.О. Сидоров *(sidorov@cic.ipn.mx)*

Н. Кастро-Санчес

Лаборатория естественного языка и обработки текста,
Центр Компьютерных Исследований (CIC),
Национальный Политехнический Институт (IPN), г. Мехико, Мексика

В статье описывается система, предназначенная для использования психологом в процессе анализа особого типа текстов – текстов эмоционального самоанализа. На основе их лингвистического анализа психолог может делать выводы об эмоциональном состоянии испытуемого или типе личности. Система предназначена для облегчения работы психолога. В ней проводится автоматический морфологический анализ, подсчитываются различные статистические параметры (частоты, лексическое богатство, и др.), отдельно представляются данные по словами, имеющим эмоциональную окраску, поскольку именно такие слова характеризуют состояние испытуемого. Реализован механизм синхронизации изменения температуры тела в момент написания текста с самим текстом. Также описано приложение системы в другой области – к анализу политического дискурса в Мексике.

Введение

Одна из основных задач прикладной лингвистики состоит в создании прикладных лингвистических систем, т. е. систем ориентированных на определенную область знаний, в которые интегрирована какая-либо лингвистическая информация. Эта лингвистическая информация оказывается полезной и используется при решении конкретных задач из этих областей.

Одной из таких областей применения лингвистических знаний и методов обработки текстов может быть психология. Более кокретно, та ее часть, которая делает выводы о психологическом состоянии или типе личности на основе анализа текстов.

Например, в 7 утвержается, что существует зависимость между частотой употребления слов служебных частей речи (предлогов, союзов, местоимений, артиклей, и др.) и различными демографическими показателями, характериситками личности и физическим и психологическим здоровьем.

Еще один характерный пример состоит в различиях в употреблении автореференции в поэтических текстах, написанных поэтами, совершившими самоубийство, в сравнении с поэтами, которые не совершили этот шаг 11. Утвержается, что совершившие самоубийство поэты чаще используют отсылки на себя (обычно, местоимения первого лица или глаголы в первом лице), и реже отсылки на других.

Еще одна интересная закономерность связана с выявлением намерения солгать или утаить информацию 9. Выявлено, что при попытке солгать в тексте на подсознательном уровне происходит следующее: используется меньше самооценивающих фраз, чаще употребляются слова с негативной эмоциональной оценкой, используется меньше когнитивно сложных маркеров.

В данной статье приводится описание системы, в которую интегрирована лингвистическая информация, для анализа особого типа текстов – текстов эмоционального самоанализа (5, 6, 7, 8). Приводится описание системы, которая реализована на испанском материале. Система является полезной во многих областях. Рассматривается ее приложение в другой предметной области – для анализа политического дискурса на основании текстов предвыборных речей кандидатов на пост президента Мексики.

Тексты эмоционального самоанализа

Написание текстов, в которых излагается или моделируется в положительной перспективе опыт, пережитый людьми в стрессовых ситуациях, помогает преодолеть отрицательные последствия этих событий 5.

Такая техника психологического анализа и лечения, когда пациент пишет по определенному плану под присмотром психолога, выражая свои мысли и чувства относительно травмирующих событий в прошлом,

[1] Работа выполнена при частичной поддержке правительства Мексики (КОНАЦиТ, СНИ) и Национального Политехнического Института (Мексика). The work was done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute (SIP, COFAA, PIFI), Mexico.

получила название техники текстов эмоционального самоанализа. Эта техника разработана на психологическом факультете Автономного национального университете Мексики совместно с некоторыми психологами из США.

Техника предполагает последующий тщательный лингвистический анализ текста психологом. До того, как система начала использоваться в этом университете, весь анализ, связанный с подсчетом статистик слов, проводился вручную. В системе предполагается как анализ содержания, т.е., что говорится, так и анализ стиля, т.е., как говорится.

Используемые лингвистические характеристики

Подсчитываются лингвистические характеристики на нескольких уровнях – стандартные статистики для текста, его морфологическая и синтаксическая структура, и семантический анализ, связанный с употреблением слов из заранее заданных списков (положительные, отрицательные, и пр.).

В качестве стандартных статистик используются следующие: количество употребленных лемм, количество употребленных словоформ, количество предложений и абзацев, средняя длина предложения и абзаца, процент употребленных вульгаризмов, лексическое богатство текста (в нашем случае мы пользовались двумя формулами –*Honoré* и *Brunét*).

Почти все характеристики являются достаточно очевидными. Для подсчета вульгаризмов исспользуется безе данных этих слов.

Поясним немного понятие лексического богатства. Заметим, что просто подсчитывать количество разных лексем в тексте не очень правильно, потому что это нелинейно зависит от длины текста, что является так называемым законом Хипса, см., например, 2.

Индекс *Brunét* вычисляется в соответствии с формулой

$$W = N^{V^{(-0.165)}}$$

где N это длина текста в словах, а V это количество использованных лексем. Полученные значения обычно находятся в интервале от 10 до 20. Чем **меньше** значение, тем больше лексическое богатство.

Другой параметр это статистика *Honoré*. Она основана на идее, что лексическое богатство в целом пропорционально количеству лексем, употребленных один раз, т.е., имеющих частоту равную единице. Эта статиситка подсчиывается по следующей формуле:

$$R = \frac{100 * \log N}{1 - (V_1 / V)}$$

где N это длина текста в словах, V это количество всех использованных лексем, и V_1 это количество лексем с частотой единица.

В данном случае, большее значение выражения соответствует большему лексическому богатству.

Для подсчета этих статистик и для любого последующего анализа важно иметь в системе морфологический анализатор. В данном случае мы пользовались разработанным в нашей лаборатории анализатором для испанского языка 1, 3 (также есть версия для русского языка, доступная для бесплатного использования).

Кроме собственно лемматизации, такой анализатор дает возможность подсчитывать статистики для употребления грамматических форм, например, форм первого или второго лица, и т. п.

Для разрешения омонимии частей речи и грамматических форм мы пользуемся частью пакета SVMTool, реализующего эту функцию. Пакет поставляется с данными для испанского языка. Он основан на формализме Support Vector Machines. Заявлена точность до 96%.

С точки зрения статистик, связанных с синтаксическими структурами, пока что подсчитывается только количество различных типов сочинительных и подчинительных конструкций.

В семантическом анализе, для целей нашей системы релевантен фактор оценки 10, связанный со шкалой *положительный-отрицательный*. Соответствующие слова были тщательно отобраны в экспериментах с участием психологов, специалистов по социальной адаптации.

Например, используются следующие эмоциональные слова и производные от них:

Слова с положительной оценкой	**Слова с отрицательной оценкой**	
	Физическая угроза	**Социальная угроза**
Откровенный	задыхаться	застенчивость
Честный	удушать	неудача
Радость	терять сознание	отвержение
Любезный	инфаркт	оскорбление
Воодушевленный	нападение, приступ	высокомерие
Удовольствие	самоубийство	бесполезность
Спокойствие	болезнь	неловкость
содержать...	сердце...	стыд...

Табл.1. Фрагмент списка эмоциональных слов

Описание системы

Система состоит из базы данных, позволяющей хранить данные о пациентах, их визитах и соответствующих текстах. Тексты могут группироваться в пользовательские корпуса или обрабатываться по отдельности.

Результаты для ручного анализа представляются в виде статистик. Кроме того, есть возможность доступа к конкретным вхождениям заданных слов в тексты для анализа контекстов их употребления.

Предоставлена возможность редактировать словари положительных и отрицательных слов и вульгаризмов.

В системе прдусмотрена возможность синхронизировать с текстом файл, содержащий измерения температуры тела, полученной в момент создания каждой части текста. Таким образом, психолог может, выбрав какой-либо фрагмент текста, видеть на диаграмме температуры, происходили ли с ней какие-либо изменения. Или наоборот, выбрав фрагмент на диаграмме температуры, где произошло ее изменение, получить доступ к соответствующему фрагменту текста и проанализировать, что могло бы вызвать такое изменение.

Применение системы в другой области (политический дискурс)

Система является достаточно универсальным инструментом лингвистического анализа. Мы также применили ее для анализа политического дискурса, а именно, предвыборных речей, произнесенных кандидатами на пост президента Мексики. Правда, очевидно, что, скажем, данные об изменении температуры при этом являются недоступными.

Любопытно, что есть характеристики, по которым Мексика очень похожа на Россию. В экономическом плане: экономика засисит от нефти, исключительная бюрократизированность государства, наличие коррупции. В социальном плане: в стране очень много бедных, существует большое расслоение общества, много национальных меньшинств (более 100). В историческом плане: революция против богатых в 1917 году, последующая гражданская война, власть одной партии в течении 71 года (до 2000 года), подавление инакомыслия в 1960-1980 годах, и, в некотором смысле, застой, вооруженная партизанская борьба в одном из штатов за независимость (хотя и несколько ограниченную, но основанную на национальной идее), идущая с 1994 года.

В данный момент в Мексике три крупные партии, которые могут выиграть выборы. Можно условно назвать их «правые» (PAN), «левые» (PRD) и центристы (PRI). Для «правых» и центристов у нас был доступ к текстам избирательной кампании 2000 года, речи же «левого» кандидата взяты из кампании 2006 года. Всего было проанализировано 73 текста (41, 16 и 16 соответственно).

Приведем некоторые полученные данные.

	«Правый»	**Центрист**	**«Левый»**
Всего слов	53,571	65,000	29,720
Словоформы	20,926	16,434	9,956
Лексемы	17,478	12,213	7,956
Лексическое богатство по *Honoré*	472.9	481.3	447.3
Лексическое богатство по *Brunét*	8.24	9.334	9.535

Табл.2. Статистика для предвыборных дискурсов

Как видно, наименьшее лексическое богатство у «левого» кандидата. Это вполне объясняется его ориентацией на самые бедные слои населения с низким культурным уровнем. Статистики лексического богатства дали противоречащие значения для «правого» кандидата и центриста, хотя различия не очень велики. Возможно, что это объясняется тем, что один из кандидатов охватил больше тем в своих выступления, и тем самым, имел возможность употребить больше слов с единичной частотой в своих выступлениях, что влияет на лексическое богатство по *Honoré*.

Оценка	**«Правый»**	**Центрист**	**«Левый»**
Положительная	уверенность (6) безопасность (4) благодарить (1)	безопасность (9) честность (8) спокойствие (4)	благодарить (3) доверие (3) безопасность (3)
Социальная угроза	оскорбить (4) критика (1) презрение (1)	стыдиться (1) ярость (1) критика (1)	отвергнутый (6) критика (1) неудача (1)
Физическая угроза	несчастный случай (1) атака (1) болезнь (1)	нападать (6) болезнь (3) рана (3)	атака (1)

Табл.3. Использование некоторых эмоциональных слов в предвыборных дискурсах

Как видно, «левый» кандидат употребляет меньше слов с положительной оценкой, больше слов с отрицательной социальной оценкой и избегает слов, связанных с отрицательной физической оценкой.

С другой стороны, центрист употребляет больше всех слов с положительной оценкой и с отрицательной физической оценкой. Эти данные носят предварительный характер и заслуживают более тщательного анализа. Мы их приводим в качестве иллюстрации возможностей разработанной системы.

Выводы

В статье была описана система, предназначенная для использования психологом в процессе анализа особого типа текстов – текстов эмоционального самоанализа. Эти тексты пишутся испытуемыми о каких-либо стрессовых ситуациях своей жизни, обычно жертвами каких-либо преступлений. На основе их лингвистического анализа психолог может делать выводы об эмоциональном состоянии испытуемого. Система предназначена для облегчения работы психолога. Она разработана для испанского языка. В ней проводится автоматический морфологический анализ, подсчитываются различные статистические параметры (частоты, лексическое богатство, и др.), отдельно даются данные по словами, имеющим эмоциональную окраску, поскольку именно такие слова характеризуют состояние испытуемого. Реализован механизм синхронизации изменения температуры тела в момент написания текста с самим текстом.

Система является достаточно универсальным инструментом лингвистического анализа. Эта же система легко была применена к анализу политического дискурса, а именно, предвыборных речей, произнесенных кандидатами на пост президента Мексики.

Литература

1. *Гельбух, А.Ф., Г.О. Сидоров*. К вопросу об автоматическом морфологическом анализе флективных языков // Труды межд. конференции Диалог-2005, М., 2005, стр. 92-96.
2. *Gelbukh, A. and G. Sidorov*. Zipf and Heaps Laws' Coefficients Depend on Language // Lecture Notes in Computer Science N 2004, 2001, Springer-Verlag, pp. 330–333.
3. *Gelbukh, A. and G. Sidorov.* Approach to construction of automatic morphological analysis systems for inflective languages with little effort // Lecture Notes in Computer Science, N 2588, Springer-Verlag, 2003, pp. 215–220.
4. *Gelbukh, A, G. Sidorov, SangYong Han.* On Some Optimization Heuristics for Lesk-Like WSD Algorithms // Lecture Notes in Computer Science, N 3513, Springer-Verlag, 2005, pp. 402–405.
5. *Domínguez, B., J. Pennebaker, y Y. Olvera.* Procedimientos no invasivos para la revelación emocional. Diseño, ejecución y evaluación de la escritura emocional autorreflexiva // 2003.
6. *Baños, R. M., S. Quero y C. Botella.* Sesgos atencionales en la fobia social medidos mediante dos formatos de la tarea de Stroop emocional (de tarjetas y computarizado) y papel mediador de distintas variables clínicas // International Journal of Clinical and Health Psychology. ISSN 1697-2600. Vol. 5, no. 1, pp. 23-42. 2005.
7. *Pennebaker, J., M. Mehl and K. Niederhoffer*. Psychological Aspects of Natural Language Use: Our Words, Our Selves // Annual Reviews Psychology, 2003.
8. *González, L., et al.* El impacto psicofisiológico y cognoscitivo de la expresión emocional autorreflexiva sobre la salud // UNAM, México, 2004
9. *Newman, M., et al.* Lying Words: Predicting Deception From Linguistic Styles // Society for Personality and Social Psychology, vol. 29, No. 5, 2003, pp. 665-675.
10. *Turney, P. D. and M. L. Littman.* Measuring Praise and Criticism: Inference of Semantic Orientation from Association // ACM Transactions on Information Systems, vol. 21, no. 4, pp. 315-346. 2003.
11. *Stirman, W. and J. Pennebaker*. Word use in the poetry of suicidal and non-suicidal poets // Psychosomatic Medicine, No. 63, 2001, pp. 517-522.