Е.В. Ягунова, Л.М. Пивоварова (СПбГУ)

Коллокации и конструкции в исследовании структуры текста

Мы рассматриваем сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и/или экспериментов с информантами.

Рассматриваемые нами сочетания (структурные составляющие текста) представляют собой неоднородное множество: с точки зрения соотнесенности со словарем и/или грамматикой, номинативностью и/или предикативностью. Типовые или ядерные **коллокации** и **конструкции** часто могут оказаться противопоставленными как парадигматические vs. синтагматические единицы (или единицы, принадлежащие лексикону vs. синтаксису).

Главным для нас является опора на следующие виды **контекста**:

*минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;

*текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;

*контекст, предполагающий учет текстов определенного типа

Вычислительный эксперимент:

Нами использовалась свободно распространяемая программа cosegment (http://donelaitis.vdu.lt/~vidas/tools.htm)

Видоизмененная мера Дайса:

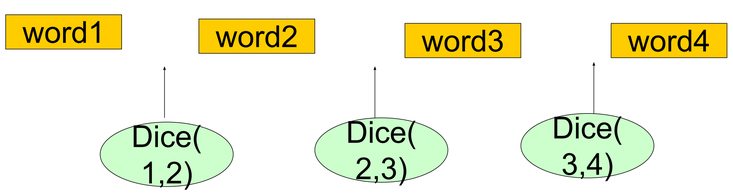
$$Dice'(x,y) = \log_2\left(\frac{2 * f(x,y)}{f(x) + f(y)}\right)$$

,где f(x) и f(y) – частота встречаемости слов x и y в коллекции, а f(x,y) – частота совместной встречаемости слов x и y.

Алгоритм:

-для всех пар слов по всей коллекции считается коэффициент Дайса

-для каждого конкретного текста «сборка» связанных сегментов:



word2 объединяется с word3 в том случае, если Dice(2,3) > [Dice(1,2)+Dice(3,4)] / 2

Таким образом получаются **цепочки слов произвольной длины**.

Эксперимент с информантами:

Информантами оценивают связность между (пробельными) словами в шкале от 0 до 5, где 5 – соответствует максимальной, а 0 – минимальной степени связности, у них карт-бланш: им не даются никакие пояснения о том, что надо понимать под связностью. Затем считается среднее арифметическое по всем информантам, два слова считаются связанными если мера связности на шкале больше или равна, чем 3,7

Предварительные результаты:

- с увеличением степени однородности (коллекция→ однородная коллекция→текст) увеличивается объем n-грамм (увеличивается n);
- с увеличением степени однородности (коллекция→ однородная коллекция→текст) увеличивается число конструкций (в соотношении конструкция vs. типовая коллокация), увеличивается число предикативных сочетаний;
- набор связанных сочетаний, подсчитанных для каждого текста отдельно в ходе вычислительного эксперимента, сходен с набором сочетаний, полученных в ходе экспериментов с информантами,
- в ходе экспериментов с информантами выделяется несколько больше предикативных сочетаний, чем в ходе вычислительного эксперимента.

Связанные сегменты, состоящие не менее чем из трех текстоформ (значимая информация, вероятные «фигуры»)

Вычислительный эксперимент			**Эксперимент с информантами**, единичный текст про А. Шварцнеггера
Коллекция (Лента.ру 2010 г)	Кластер про Шварцнеггера (однородная коллекция)	Единичный текст про А. Шварцнеггера	
тем не менее	глобальное инновационное партнерство	только что приземлился	Губернатор Калифорнии Арнольд Шварценеггер
в связи с	представителей ведущих компаний	**могу дождаться встречи**	прилетел в Москву.
в 2009 году	с губернатором калифорнии	вскоре после этого	в российскую столицу
то же время	могу дождаться встречи	**ответил** калифорнийскому губернатору	Не **могу дождаться встречи** с президентом Медведевым
в настоящее время	во главе делегации	англоязычная версия твита	российский президент Дмитрий Медведев **ответил**
со ссылкой на	создать настоящий технологический бум	ответил ему взаимностью	в своем микроблоге
возбуждено уголовное дело	сфере высоких технологий	это же время	добро пожаловать в Москву
по сравнению с	только что приземлился		Жду встречи с вами
в 2008 году	тогда вам сказал		Медведев добавил микроблог
			с делегацией представителей
Полужирный шрифт: сегменты или их фрагменты, присутствующие в обоих списках (3 и 4 графа). В графу 2 попала верхушка наиболее частотных связанных сегментов, упорядоченных по частоте, остальные графы представлены в полном объеме.			он встретится с российскими министрами
			во время посещения Медведевым
			российский президент завел себе

Структура текста по данным информантов (см. графу 4). П/ж шрифтом выделены фигуры

Губернатор Калифорнии Арнольд Шварценеггер 10 октября **прилетел в Москву.** / После прибытия **в российскую столицу** он сделал в своем микроблоге на Twitter соответствующую запись (Только что приземлился в Москве. Прекрасный день. **Не могу дождаться встречи с президентом Медведевым**), а также разместил фотографию, сделанную по дороге из аэропорта.

Вскоре после этого **российский президент Дмитрий Медведев ответил** калифорнийскому губернатору **в своем микроблоге**: @Schwarzenegger, **добро пожаловать в Москву**. Англоязычная версия твита Медведева также содержала слова "**Жду встречи с вами** и вашей делегацией в @skolkovo".

Кроме того, **Медведев добавил микроблог** Шварценеггера в друзья. Губернатор Калифорнии ответил ему взаимностью.

Как сообщает РИА Новости, Шварценеггер приехал в Россию **с делегацией представителей** венчурных фондов и инновационных компаний Кремниевой долины. Планируется, что помимо президента Медведева, **он встретится с российскими министрами**.

Президент России и губернатор Калифорнии **в этом году** уже встречались - **это произошло в июне / во время посещения Медведевым** США. В это же время **российский президент завел себе** микроблог.