С.В. Зинин

Проект «Сражающиеся Царства»
Массачусетский университет (Амхерст)

Корпусный анализ семантических отношений иероглифов и ключей

Китайские иероглифы могут быть разбиты на семантические группы, в соответствии с принятой в китайской орфографии системой ключей. Структура большинства китайских иероглифов включает два компонента: ключ (детерминатив) и фонетик. Фонетик определяет (приблизительное) произношение иероглифа, а ключ задаёт широкую область его значения. Ключи представляют собой традиционное и удобное средство поиска и организации иероглифов. В современной китайской орфографии принят список из 214 ключей Канси. В соответствии со списком ключей, иероглифы образуют группы, семантика которых, предположительно, может описываться значением ключа (группы ключей)[1].

Идея организации иероглифов в соответствии с ключами близка современной идее семантической онтологии (понятиям высшего уровня, top ontology). Традиционное деление иероглифов на семантические группы по ключам привлекает в последнее время внимание специалистов в области вычислительной лингвистики, занимающимися проблемами онтологий (классический пример такой онтологии – WordNet). Некоторые из китайских лингвистов предлагают принять в качестве такой онтологии для китайского языка не национальный вариант онтологии WordNet (например, HowNet), а систему ключей Канси[2].

Использование системы Канси в качестве онтологии вполне возможно, так как между большинством иероглифов с одинаковым ключом практически всегда можно найти какие-то смысловые сходства. Но построение семантических онтологий – достаточно субъективный процесс. Возможны ли какие-то объективные критерии, с помощью которых можно было бы обосновать выбор той или иной онтологии[3]? Ответ на этот вопрос требует выполнения исследования, и поэтому интересны любые объективные данные о семантических отношениях иероглифов и, в особенности, степени семантической близости иероглифов в группах, образованных на

основании ключей. Если бы удалось с помощью методов вычислительной лингвистики показать, что между иероглифами с одним и тем же ключом существуют достаточно сильные семантические связи, то это было бы сильным аргументом в пользу выбора системы Канси как семантической онтологии.

Например, методы матричной декомпозиции (SVD) позволяют проводить кластерный анализ термов корпуса, а методы латентно-семантического анализа (LSA/LDA) позволяют выделить тематические группы семантически связанных термов (иероглифов) в корпусе. Если в выделенных тематических группах значительное количество иероглифов имело бы тот же ключ, то это означало бы, что семантические связи, описываемые ключами, достаточно сильны[4] (см. [12; 14]).

С помощью методов SVD можно оценить семантические расстояния между иероглифами. Если представить себе, что группы иероглифов, образованные с помощью ключей, представляют собой семантически значимые образования, то семантически иероглифы группы будут ближе друг к другу, чем к другим группам.

Для проведения такого рода корпусных исследований необходимо иметь достаточно большой корпус. В современных условиях найти такой корпус можно только для современного китайского языка. Однако слова современного китайского языка в массе своей двусложны, то есть, состоят из двух иероглифов [13]. Насколько значим будет анализ такого корпуса для понимания семантических связей отдельных иероглифов?

Мы провели с этой целью тематический анализ корпуса современного китайского языка университета Лидса, а также сравнительно небольшого корпуса классических текстов (слова в которых можно условно считать односложными) [1; 15]. Анализ проводился с применением нескольких параметров, а также различных выборок текстов. Было получено несколько интересных результатов[5].

Во-первых, мы установили, что тематические группы для двусложных слов (современного языка), и для иероглифов весьма близки. Это позволяет утверждать, что представление современных (в основном, двухсложных) текстов как текстов, состоящих из отдельных иероглифов, достаточно репрезентативно семантически с точки зрения вычислительной лингвистики.

Во-вторых, мы установили, что тематические группы иероглифов действительно содержат иероглифы, которые можно объединить в рамках определённой темы. То есть, методы латентно-семантического анализа вполне приложимы к корпусам современного и классического китайских языков, в том числе, представленным как отдельные иероглифы. Однако эти методы не могут представить основания для определения ключей как организаторов семантических кластеров. Иероглифы с одним и тем же ключом не составляют там значительную долю.

Последний вывод следует понимать как аргумент в пользу неприложимости корпусных методов латентно-семантического анализа для обоснования системы ключей Канси как семантической онтологии.

Методы кластерного анализа позволяют проанализировать группы ключей как кластеры. Однако, в нашем случае, кластеры демонстрируют более сильные семантические связи между иероглифами, не имеющими прямого отношения к ключам группы. Иначе говоря, хотя ключи и определяют некоторые семантические отношения между иероглифами с одним и тем же ключом, эти отношения могут быть слабее, чем семантические отношения между иероглифами с разными ключами[6].

В тех случаях, когда ключи задают наиболее сильные семантические отношения между иероглифами, речь идёт о сравнительно небольших группах в 3–5 иероглифов. Это может служить аргументом в пользу того, что ключевые группы, куда входит более 5 иероглифов, следует дробить на более мелкие группы. Иначе говоря, на поле ключей должна быть создана более сложная, многоуровневая структура (как это и было в системе ключей, первоначально предложенной Сюй Шэнем, см. [3]).

В целом, хотя корпусный подход и не позволил подтвердить гипотезы об тесных семантических связях между иероглифами в группах ключей, он, несомненно, позволяет понять, как лучше использовать корпусные технологии для сематического анализа китайских иероглифов.

Примечания

[1] См. историю возникновения системы ключей в [2; 3].

[2] См. напр., [4; 5; 6; 8].

[3] Например, что лучше – система ключей Канси или национальный вариант WordNet? См. напр., [16; 17].

[4] Разумеется, мы не ожидаем, что выделенные методами вычислительной лингвистики тематические группы будут однозначно соответствовать группам ключей. Тем не менее, следовало бы ожидать, что в некоторых из таких групп, иероглифы с одинаковым ключом будут составлять значительную долю (например, иероглифы с ключом «человек» в топике «общество»).

[5] В данном сообщении мы опускаем количественные данные эксперимента, а также списки иероглифов в тематических группах, данные кластерного анализа для групп ключей, и так далее. Они будут включены в расширенное сообщение.

[6] Для сравнительного анализа, мы также проводили эксперименты с кластерами, образованными слогами языка. Они показали, в целом, несколько меньшую семантическую «связность», чем ключевые кластеры.

Литература

1. CTEXTS: http://www.umass.edu/ctexts/index.php

2. *Boltz W.G*. The Origin and Early Development of the Chinese Writing System. American Oriental Series, vol. 78. New Haven, 1994.

3. *Bottéro*, *F.* Sémantisme et classification dans l'écriture chinoise, les systèmes de classement des caractères par clés du Shuowen jiezi au Kangxi zidian. Paris, Institut des Hautes Études Chinoises, Collège de France, 1996.

4. *Cai Dongfeng*, *Sun Jingguang*, *Zhang Guiping*, *Lü Dexin*, *Dong Yanju*, *Song Yan* and *Chao Yu*. HowNet Based Chinese Question Classification // Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 20). Wuhan, 2006. Pp. 366–389.

5. *Chou Yamin*, *Huang Churen*. Hantology: An Ontology based on Conventionalized Conceptualization // *Huang Churen* et al. (Eds.) Ontologies and Lexical Resources for Natural Language Processing. Cambridge: Cambridge University Press, 2008.

6. *Chou Yamin*, *Hsieh Shukai* and *Huang Churen*. Hanzi Grid: Toward a Knowledge Infrastructure for Chinese Character Based Cultures // *Ishida T., Fussell S.R., Vossen P.T.J.M*. (eds.). Intercultural Collaboration I. Lecture Notes in Computer Science. Heidelberg: SpringerVerlag. 2007.

7. *Galambos I*. Orthography of early Chinese writing. Budapest: Eötvös Loránd University, Department of East Asian Studies. 2006.

8. *Hsieh Shukai*. Hanzi, Concept and Computation: A Preliminary Survey of Chinese Characters as a Knowledge Resource in NLP // Philosophische Dissertation angenommen von der Neuphilologischen Fakultät der Universität Tübingen. 2006.

9. *Hu He, Du Xiaoyong*, *Tian Xuan* and *Bai Ruixue*. A Preliminary Study on the Semantic Strength of Chinese Radicals. // Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007). Vol. 2. 2007. Pp. 658–662.

10. *Hu He* and *Du Xiaoyong*. A semantic analysis of Chinese radicals // International Journal of Business Intelligence and Data Mining. 2008, #3(4). Pp. 426–436.

11. *Huang Churen*, *Yang Yajun* and *Chen Shengyi*. An Ontology of Chinese Radicals: Concept Derivation and Knowledge Representation based on the Semantic Symbols of Four Hoofed-Mammals. Presented at The 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC2008). Philippines: De La Salle University-Manila, 2008. Pp. 189–196.

12. *Landauer T.K.*, *Foltz P.W.* and *Laham D.* Introduction to Latent Semantic Analysis // Discourse Processes, 1998. #25, Pp. 259–284.

13. *Packard*, *J.L.* The Morphology of Chinese: A linguistic and cognitive approach. Cambridge, Cambridge University Press, 2000.

14. *Řehůřek*, *R*. Gensim project: http://nlp.fi.muni.cz/projekty/gensim/index.html.

15. *Sharoff S*. Creating general-purpose corpora using automated search engine queries // Marco Baroni and Silvia Bernardini (eds.). WaCky! Working papers on the Web as Corpus. Bologna, 2006.

16. *Wong Shunha*, *S.* Fighting Arbitrariness in WordNet-like Lexical Databases –A Natural Language Motivated Remedy // *Sojka P., Pala K., Smrž P., Fellbaum C., Vossen P.* (Eds.): The Proceedings of the 2nd Global WordNet Conference (GWC 2004). Brno, 2003. Pp. 234–241.

17. *Wong Shunha*, *S.* and *Pala K*. Chinese Characters and Top Ontology in EuroWordNet. // Proceedings of the Global WordNet Conference'2002. Mysore: Mysore University, 2002. Pp. 224–233.