Designing an Affiliation Extractor for Turkish Universities through Finite State Graphs

Zehra Taşkın & Umut Al

{ztaskin, umutal}@hacettepe.edu.tr

Plan

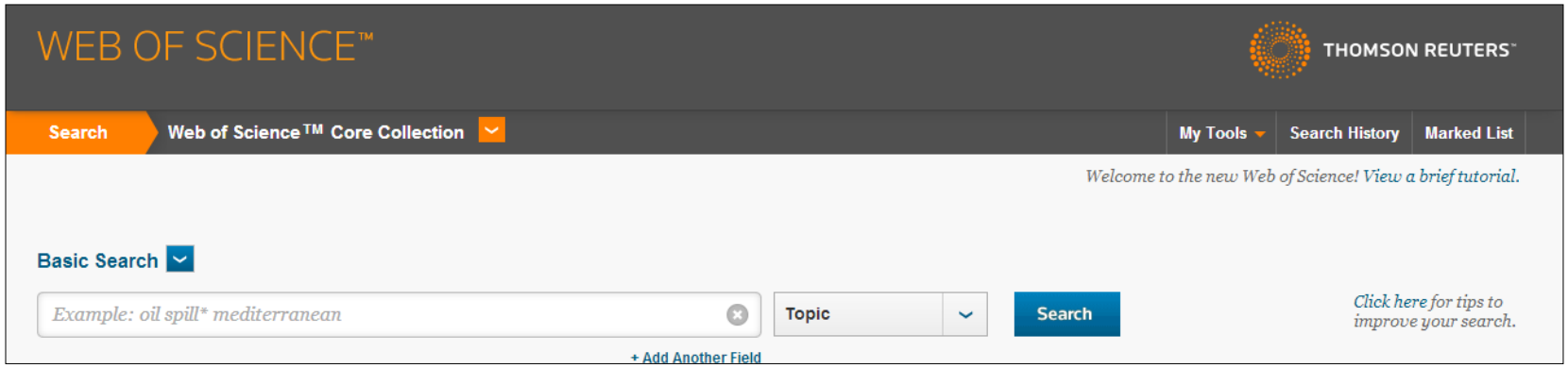
- Information retrieval and its relation to bibliometrics
- *Web of Science* and citation indexes
- Data inconsistency in citation indexes
- Methodology and the aim of the study
- Affiliation extractor model for Turkish Universities

Information Retrieval and its Relation to Bibliometrics

- Information retrieval problem (high volume natural language texts)
- Bibliometrics is the the application of mathematical and statistical methods to books and other media of communication (Pritchard, 1969, p. 348)
- Research evaluation
  - Fund distributions
  - Academic appointments and incentives
  - Impact of scientific outputs
  - Science policy making

WoS and Citation Indexes

- A platform and indexes
  - *Science Citation Index (SCI), Social Sciences Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI*)
- One of the main sources for research evaluation
- Problem: Natural language indexing



Data Inconsistency in Citation Indexes

- WYSIWYG
  - Institution names
  - Author names
  - Journal names
  - …
- Character or spelling errors
- Translation errors
- Indexing errors
- Standardization errors

Examples

- Harvard Univ => Harward Univ
- Hacettepe Univ => Hacetteppe Univ
- Univ Trakya => Univ Trakia
- Dumlupinar Univ => Durnlupinar Univ

- Standardization errors;
  - Hacettepe Hosp >> Hacettepe Univ
  - Hacettepe Fac Med >> Hacettepe Univ

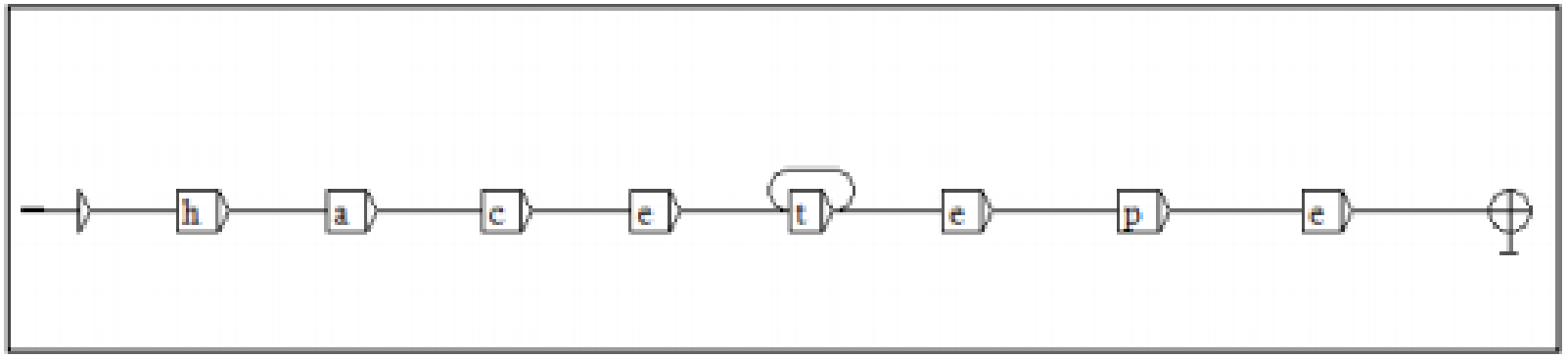
Methodology

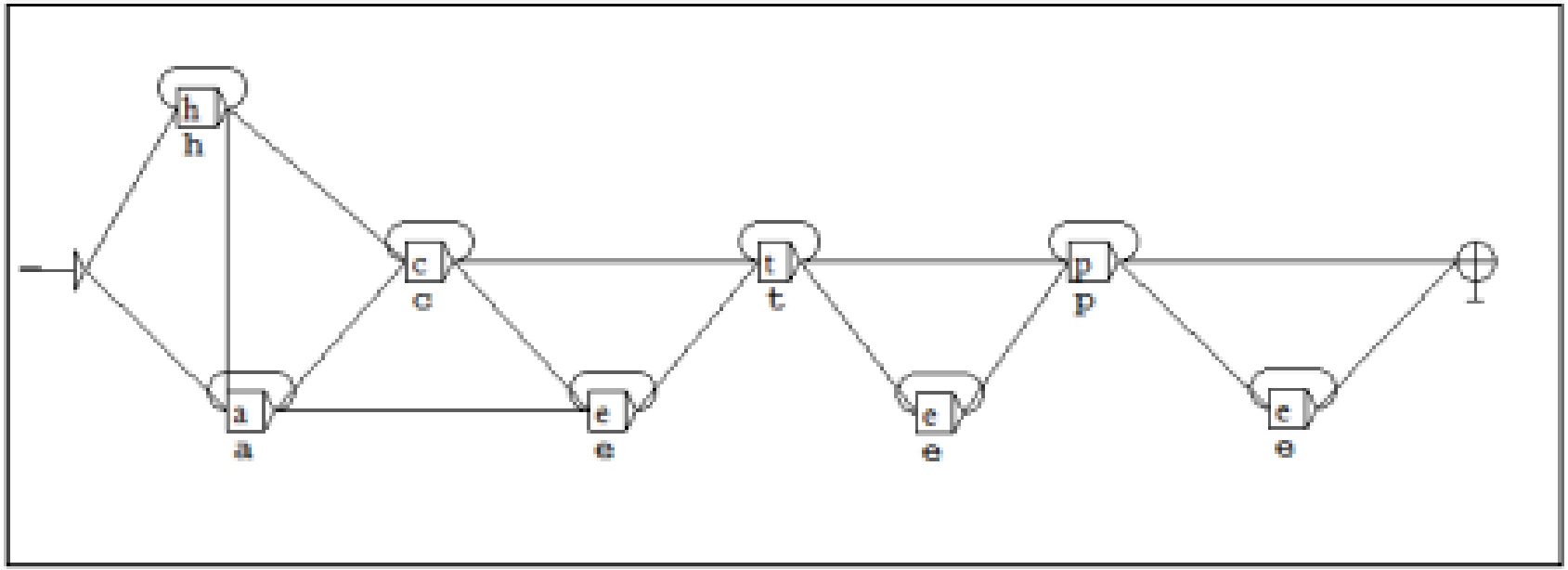
- Data source: *Web of Science*
- 197,687 Turkey-addressed publications
  - Published between 1928-2009
  - Deep data cleaning and unification process
  - The addresses of 50 universities that have more than 1,000 publications were analyzed
- Nooj for finite state graphs

Aim of the Study

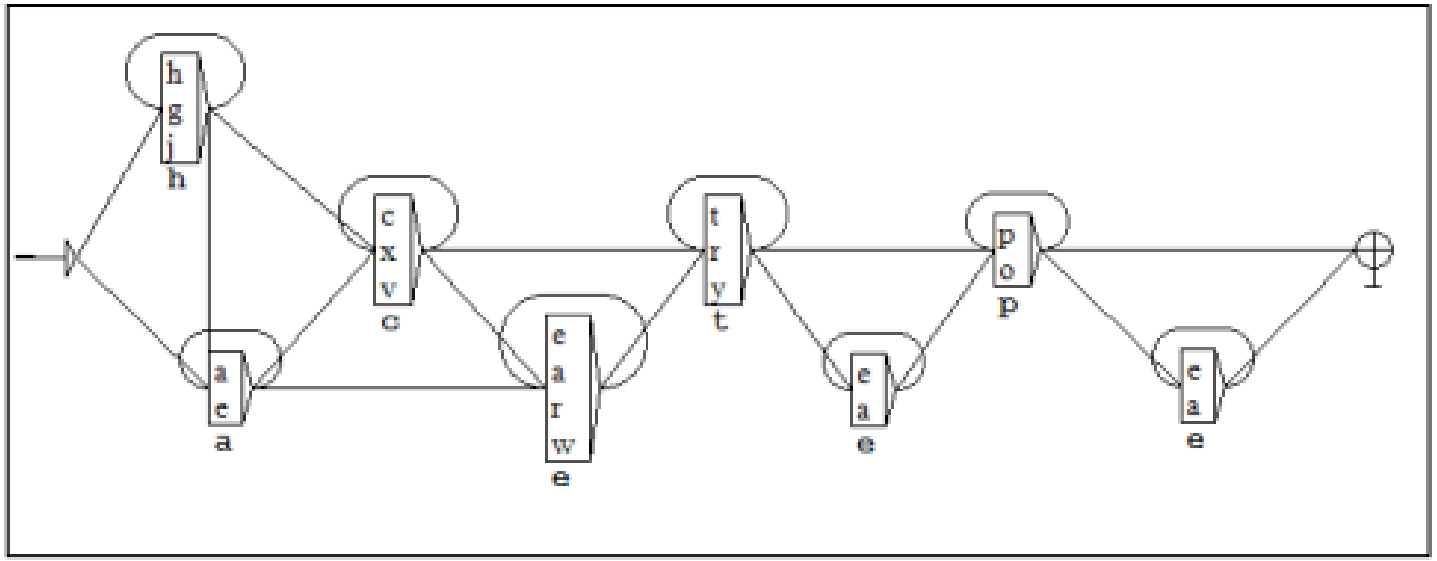
- Designing an extractor for the identification of Turkish Universities' affiliations by using finite state graphs
- Testing the possibility of employing machine learning for the task of affiliation identification and extraction by using finite state graphs

Background
(Taşkın & Al, 2014)



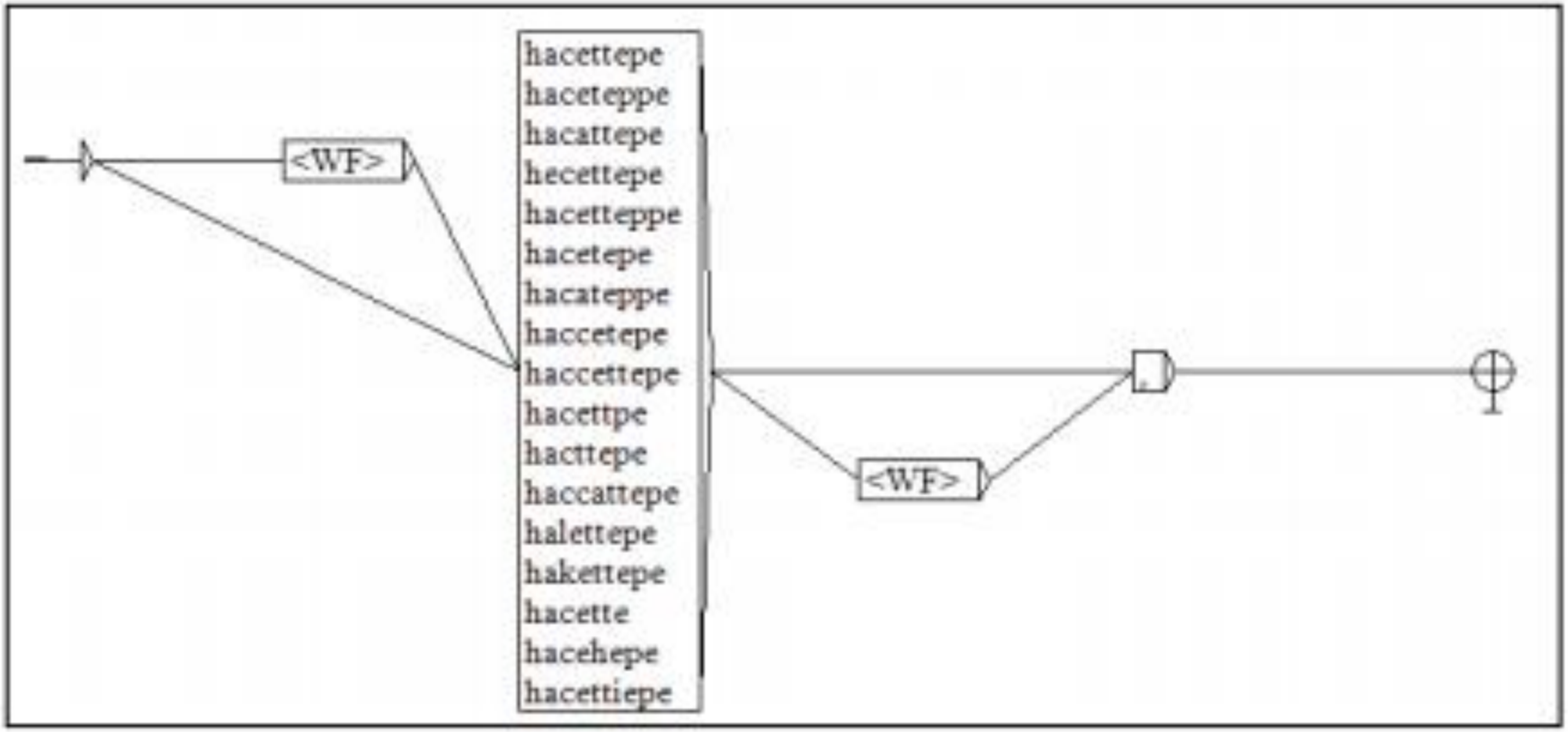


Background



Term	Freq.	Term	Freq.
Hacettepe	21,024	Hacateppe	2
HACETTEPE	4,658	HACCETEPE	2
HACETEPPE	9	HACCETTEPE	2
Hacattepe	7	HACETTPE	2
HECETTEPE	5	Hacttepe	1
HACETTEPPE	4	Haccattepe	1
Hacetepe	3	Haccettepe	1
Hecettepe	3	HACETEPE	1

Background


<WF>
hacettepe
haceteppe
hacattepe
hecettepe
hacetteppe
hacetepe
hacateppe
haccetepe
haccettepe
hacettpe
hacttepe
haccattepe
halettepe
hakettepe
hacette
hacehepe
hacettiepe
<WF>

Background

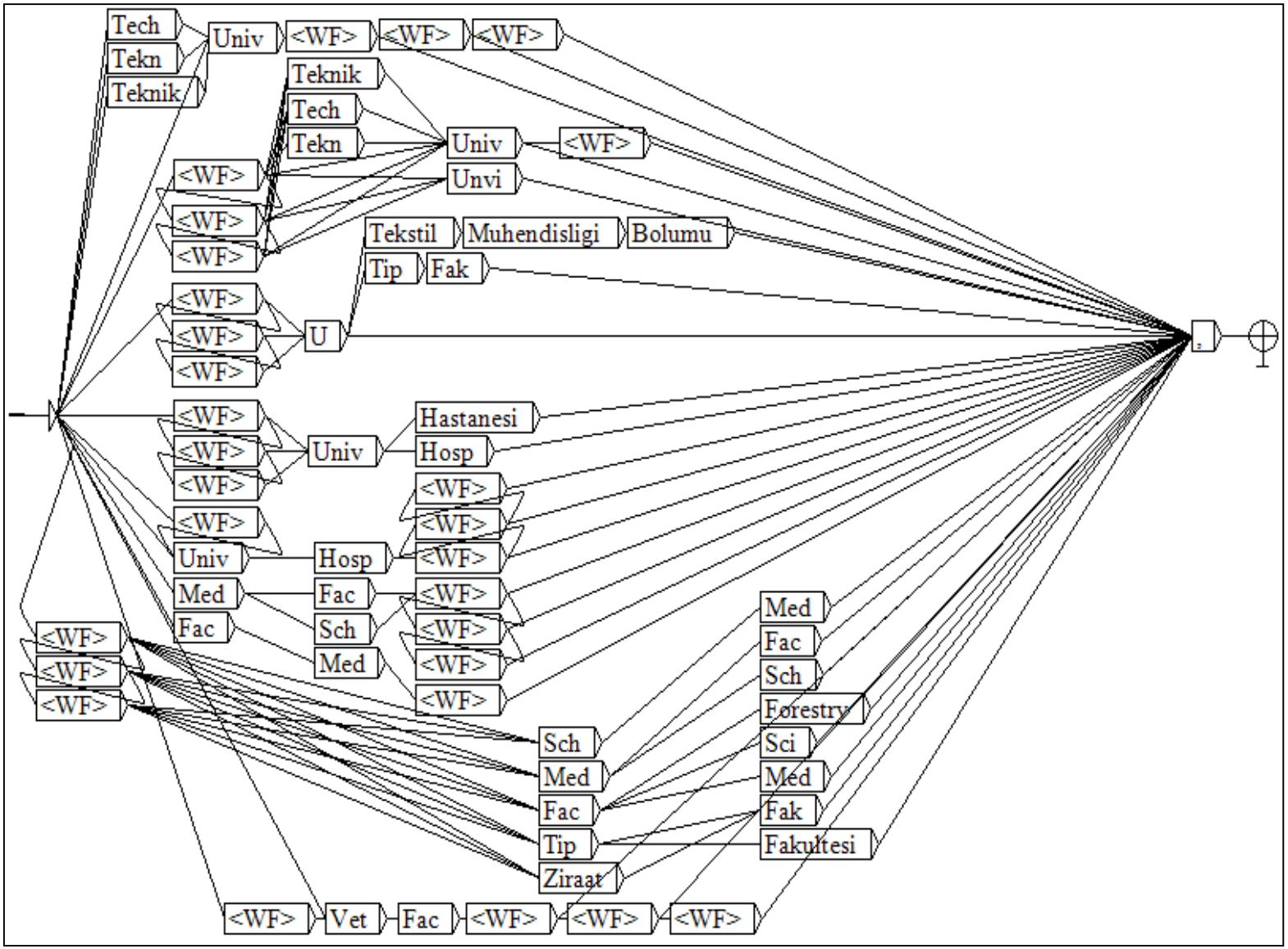
Term	Freq.	Term	Freq.
Hacettepe Univ / Univ Hacettepe	18,826	Haccetepe Fac Med	1
Hacettepe Childrens Hosp	84	Haccettepe Univ Hosp	1
Hacettepe Med Sch	60	Hacettepe Med Acad	1
Hacettepe U	33	Hacettepe Kuniv Hastaneleri	1
Hacettepe Sch Med	28	Hacehepe Univ	1
Hacettepe Med Fac	23	Haceteppe Childrens Hosp	1
Hacettepe Fac Med	12	Hacette Unıv	1
Hacettepe Hastaneleri	9	Hacettepe Technopolis	1
Hacettepe Hosp	9	Sociales Hacettepe	1
Hacettepe Hastanesi	7	Hacettepe Adult Hosp	1
Hacettepe Oncol Inst	5	Hacettepe Child Hosp	1
Haceteppe Univ	5	Hacettepe Cocuk Hastabanesi	1
Hecettepe Univ	5	Hacttepe Univ	1
Hacettepe Med Ctr	4	Hakettepe Childrens Hosp	1
Hacettepe Tip Fak	3	Halettepe Univ	1
Hacattepe Univ	3	Univ Haceteppe	1
Hacettepe Children Hosp	3	Hacettepe Cocuk Hastahanesi	1
Hecettepe Univ	3	Hacettepe Cocuk Hastanesi	1
Hacetepe Univ	2	Hacettepe Cocuk Hastenesi	1
Hacetteppe Univ	2	Hacettepe Eriskin Hastanesi	1
Hacettpe Univ	2	Hecettepe Childrens Hosp	1
Hosp Hacettepe	2	Klinikum Hacettepe	1
Hacattepe Univ Hosp	1	Hacettepe Inst Oncol	1
Haccattepe Univ	1	Unit Hacettepe	1

Findings

- A total of 433 rules for 50 universities were found

Rule	*Freq*	*% (in total)*
? Univ	130,809	64.48
? ? Univ	22,522	11.10
Univ ?	16,650	8.20
? Tech Univ	11,459	5.64
? ? Tech Univ	10,301	5.07
? ? ? Univ	2,872	1.41
Tech Univ ?	1,675	0.82
Univ ? ?	870	0.42
? Fac Med	820	0.40
? (Unidentified)	810	0.39
? Med Fac	604	0.29
? Med Sch	490	0.24
? Sch Med	245	0.12
Univ ? ? ?	199	0.09
? Tekn Univ	170	0.08
? Tip Fak	115	0.05
Fac Med ?	101	0.04
? Childrens Hosp	91	0.04
? ? Tekn Univ	84	0.04
? U Tekstil Muhendisligi Bolumu	82	0.04

The FSG Model



Concordance of Founded Affiliations

Before	Seq.	After
Chirurg, Berlin, Germany...	Univ Padua,	Dept Surg & Gastroenterol Sci, ...
[Agaoglu, Galip; Erol, O....	Istambul Bilim Univ,	ONEP Plast Surg Sci Inst
[Agaoglu, Galip; Erol, O....	Istanbul Bilim Univ,	ONEP Plast Surg Sci Inst
Turkey [Ak, Koray; Bes...	Marmara Univ Hosp,	Sch Med, Dept Cardiovasc Surg
Maher; Civelek, Ali; Ars...	Marmara Univ,	Sch Med, Dept Cardiovasc Surg
Surg, Istanbul, Turkey; [...	Abant Izzet Baysal Univ,	Dept Gen Surg, Bolu, Turkey
Surg, Bolu, Turkey; [Ka...	Istanbul Tech Univ,	Fac Mech Engn, Dept Hydromech
Sengul, Goksin; Aydin, I...	Ataturk Univ,	Sch Med, Dept Neurosurg, Erzur..
Hasan; Soyturk, Mujde; ...	Dokuz Eylul Univ,	Fac Med, Div Gastroenterol, TR
Hasanefendioglu; Sentur...	Univ Dicle,	Dept Radiol, Sch Med, TR
Turkey [Akcay, A.; Kor...	Kahramanmaras Sutcu Imam U...	Dept Cardiol, Fac Med, TR
Turkey [Akcora, Buelent...	Mustafa Kemal Univ,	Sch Med, Dept Pediat Surg
Turkey; [Akbas, Ali; Oe...	Gaziosmanpasa Univ,	Fac Med, Dept Biochem, Tokat
Tulay, C. M.; Doner, E.]	Eskisehir Osmangazi Univ,	Fac Med, Dept Thorac Surg
Kurt, Yavuz] Haydarpas...	Gulhane Mil Med Fac,	Istanbul, Turkey; [Yitgin, Selahatt..

Limitations & Future Studies

- The rule list for Turkish universities created manually due to not to lose any variations of affiliations

- This study can provide a basis for future studies focusing on automatic learning algorithms for affiliations to measure the success of machine learning

Conclusion

- This model could be extracted 99.05% of the rules
  - The affiliation extraction based on the general identification of main affiliation patterns for Turkish universities, can help the future studies
- Rule list creation is time consuming and impractical
  - However, it is more useful for the future studies that used machine learning algorithms, since it provides opportunity for comparison

References

- Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25(4), 348-349.
- Taşkın, Z. & Al, U. (2014). Standardization problem of author affiliations in citation indexes. *Scientometrics*, 98(1), 347-368.

Designing an Affiliation Extractor for Turkish Universities through Finite State Graphs

Zehra Taşkın & Umut Al

{ztaskin, umutal}@hacettepe.edu.tr