The Distiller Framework:

current state and future challenges

Marco Basaldella | Giuseppe Serra | Carlo Tasso

Artificial Intelligence Laboratory @ University of Udine - http://ailab.uniud.it

Outline

- Information Extraction in the AILAB
- The Distiller Framework
  - Motivation
  - Architecture
- Applications

KE@AILab Udine

Keyphrase Extraction

- “The automatic selection of **important, topical phrases** from within the body of a Document”; “A keyphrase list [is defined] as a short list of phrases (typically five to fifteen noun phrases) that capture the main topics discussed in a given document.” (Turney)
- The task of identifying phrases that “provide semantic metadata that **summarize and characterize documents**.” (Witten et al.)
- “the task of a keyword extraction application is to automatically identify in a text a set of terms that **best describe** the document.” (Mihalcea and Tarau)

Applications: automatic tagging, recommender systems, community analysis, social network analysis, …

Example

Text

Since 2005, the Italian Research Conference on Digital Libraries has served as an important national forum focused on digital libraries and associated technical, practical, and social issues. IRCDL encompasses the many meanings of the term "digital libraries", including new forms of information institutions; operational information systems with all manner of digital content; new means of selecting, collecting, organizing, and distributing digital content; and theoretical models of information media, including document genres and electronic publishing. Digital libraries may be viewed as a new form of information institution or as an extension of the services libraries currently provide. Representatives from academe, government, industry, and others are invited to participate in this annual conference. The conference draws from a broad and multidisciplinary array of research areas including computer science, information science, librarianship, archival science and practice, museum studies and practice, technology, social sciences, and humanities.

Keyphrases

- conference
- Research
- digital libraries

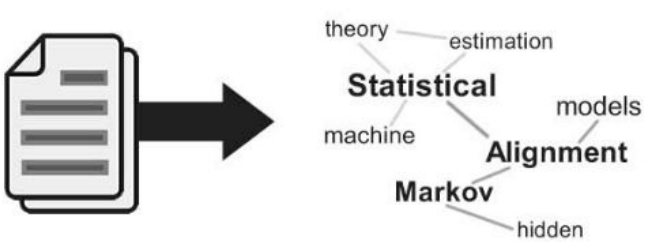
Do it yourself:

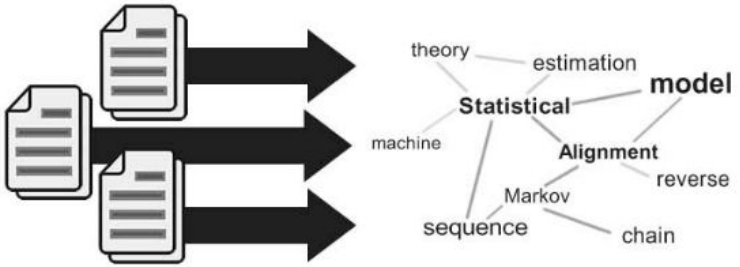
http://ailab.uniud.it/distiller

Before the Distiller

Roots in recommender systems

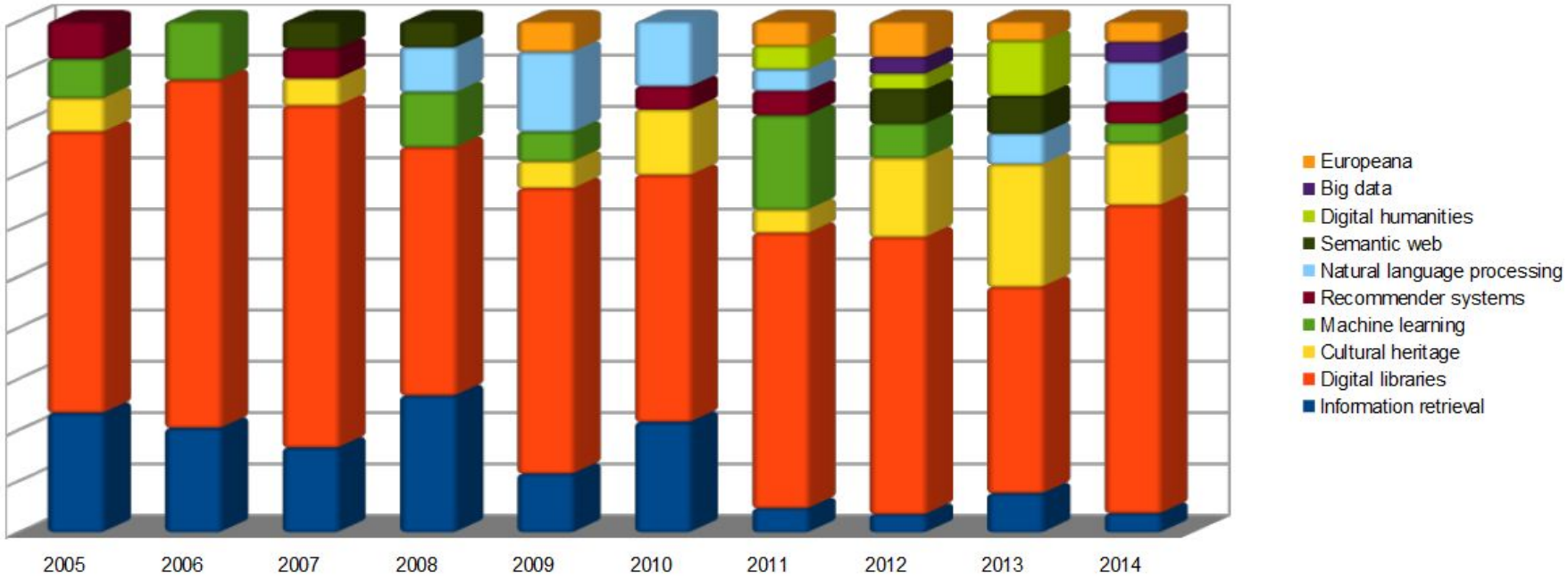
- Pudota et al., 2010: DIKpE system
  - Unsupervised Keyphrase Extraction
  - KPs as input for PIRATES recommender system
- De Nart et al. 2011-2014: applications
  - User modelling, community modelling, etc
- Degl'Innocenti, 2014: Multilinguality
  - System able to extract KPs in Italian and English





Example: KPs for community analysis

Evolutions of the topics of the IRCDL conference, 2005-2014 (De Nart et al., 2014)



The Distiller Framework

Distiller

Starting from our experience, we built Distiller as a tool with a

- **Customizable pipeline**
- **Shared components**
- **Multilanguage support**

Bonus:

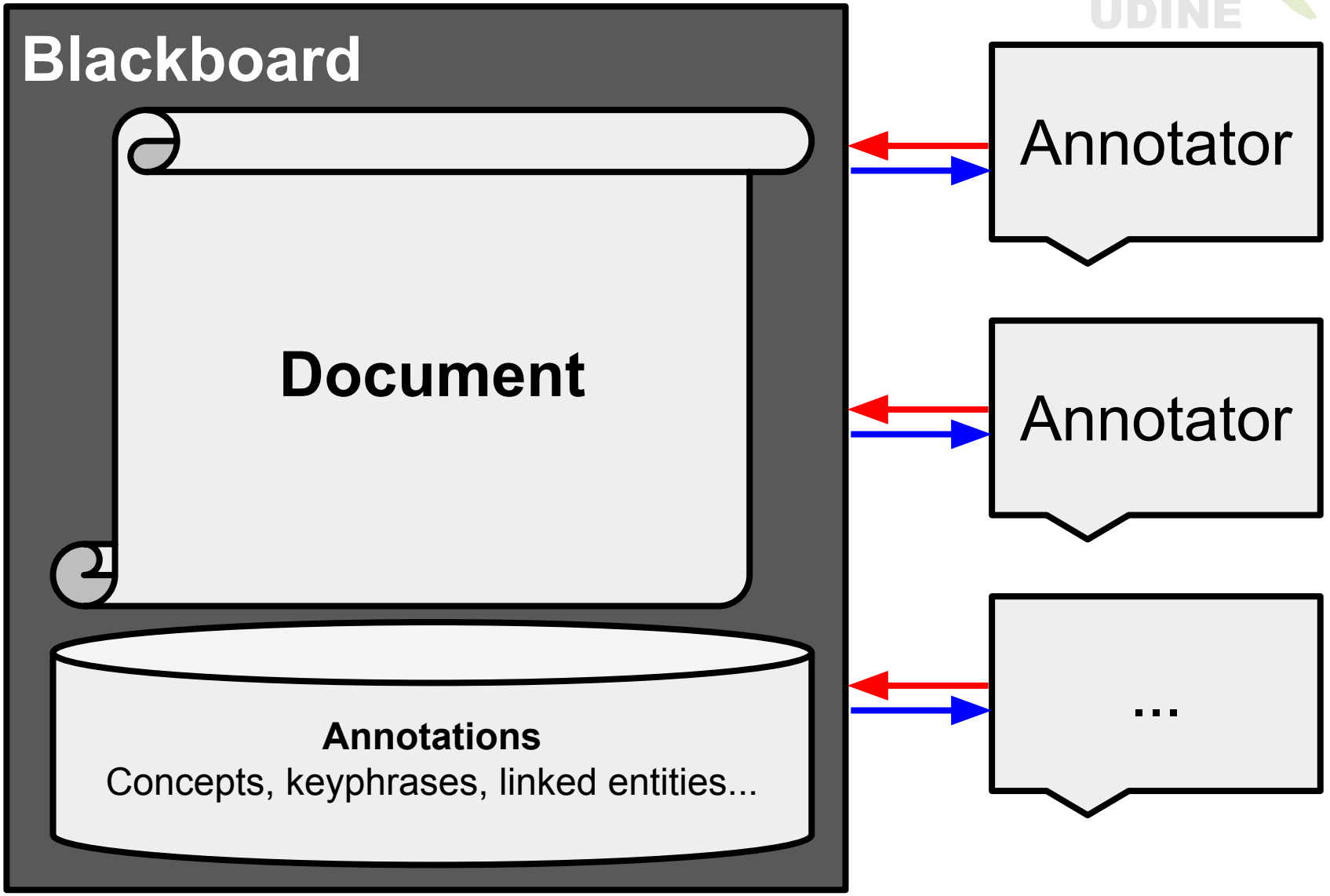
- You don’t have to reinvent the wheel: a set of “basic” components is already available
- You don't need to recompile everything to customize your pipeline (XML configuration files)

Architecture

Simple concepts:

1. **Blackboard**: holds the **document** and the **annotations**
2. **Document**: structured as a tree
3. **Annotations**: everything that can be said about the document or one of its components
4. **Annotators**: the actual “business logic”

Architecture



Architecture

Annotators are the core of the Distiller framework. They:

- Write **Annotations** on the Blackboard and on (parts of) the document
- read annotations produced by other annotators
- implement the actual algorithms (e.g. KE, NER…)
- wrap 3rd party libraries (e.g. Stanford CoreNLP)
- communicate with online 3rd party services (e.g. TagMe)

Applications

Linguistic-based AKE

We enhanced AKE by means of anaphora resolution

- Idea 1: if a concept is referenced many times in a document, it may be important
  - E.g. John Doe is a businessman. → *businessman* refers to *John Doe*

- Idea 2: substitute pronouns
  - E.g. John Doe is a businessman. He is the CEO of ACME Corporation. → *He* refers to *John Doe*. → **replace**
  - John Doe is a businessman. *John Doe* is the CEO of ACME Corporation. → now we know that **both businessman and CEO** refer to John

Basaldella, Chiaradia and Tasso (2016)

Multilingual AKE

We developed a multilingual AKE pipeline that works in five languages

- English
- Italian
- Arabic
- Portuguese
- Romanian

We demonstrated that we **don't need** training data for every language to extract KPs efficiently (Basaldella et al., 2017)

Datasets used:

- SEMEVAL 2010 for English
- We **collected** a dataset for Arabic (http://github.com/ailab-uniud/akec) (Helmy et al., 2016)

Multilingual AKE

Text

الرحلات الناجحة الى المريخ [عدل] . 1. مارينر 4 . مقالة مفصلة : مارينر 4 . مارينر 4 و مع ه المسبار مارينر 3 هما مسباران ارسلت هما ناسا الي المريخ عام 1964 و اطلق في 28 نوفمبر 1964. و كان مارينر 4 هو المسبار الرابع الذي يرسل ل اكتشاف كوكب المريخ المجاور ل الارض في المجموعة الشمسية . و كان مخططا ل ه اجراء اول مسار ب القرب من المريخ و ارسال صورا ل سطح ه الي الارض . و التقط مارينر 4 اول صور من الفضاء البعيد تعود الي الارض تبين سطح ه و قد غطت ه الفوهات , و تظهر ل نا عالما يبدو ميتا , و يغير النظرة القديمة التي كانت تعتقد ب وجود حياة علي المريخ . و قد صمم مارينر 4 ل اجراء مشاهدة عن قرب ل كوكب المريخ و ارسال الصور الي الارض . و اختصت اكتشافات ه الاخري قياس الجسيمات في تلك المناطق من الفضاء ب القرب من المريخ . ك ما كانت تلك البعثة ب غرض التعرف علي الامكانيات التكنولوجية ل السفر عبر الفضاء ل مدد زمنية طويل ه . و في 21 ديسمبر 1967 انقطع الاتصال بين مارينر 4 و الارض . 2. مارينر 6 و مارينر 7 . مقالة مفصلة : مارينر 6 و مارينر 7 . مارينر 6 و مارينر 7 هما مسباران ارسلت هما ناسا عام 1969 في اطار برنامج مارينر ل دراسة المريخ , و قد اتخذ احد هما مسارا عند خط الاستواء المريخي و الاخر مسارا قطبيا جنوبيا و قاما ب تحليل الغلاف الجوي ل المريخ و تصوير مئات من الصور ل سطح ه . و كان الغرض ايضا التمهيد التكنولوجي ل بعثات اخري الي المريخ . و قد استخدمت المعلومات التي بعث ب ها مارينر 6 ل برمجة مارينر 7 الذي خلف ه ب مدة 5 ايام . و قد اقلع و في 29 يوليو 1969 و قبل الوصول . A ب مركز كينيدي ل الفضاء و اقلع مارينر 7 من المنصة B 36 مارينر 6 من منصة الاقلاع 36 الي اقرب نقطة علي المسار ب النسبة ل المريخ ف قد الاتصال مع مارينر 7, و تبين لاحقا ان احد البطاريات في ه قد انفجرت . و رغم ذلك ف كانت المعلومات التي ارسل ها مارينر 7 الي الارض قبل حدوث عطل البطارية كان مفيدا ل الغاية حيت روعيت في اخد ها بعض بيانات مكتسبة من مارينر 6. 3. مارينر 9 . مقالة مفصلة : مارينر 9 . مسبار مارينر 9 هو مسبار ارسلت ه ناسا ل اكتشاف المريخ في اطار برنامج مارينر . و قد اطلق مارينر 9 يوم 30 مايو 1971 من كاب كانافيرال ب فلوريدا و وصل المريخ في 13 نوفمبر من نفس العام . اتخذ مارينر 9 مدارا حول المريخ , و اصبح اول قمر صناعي يتخذ مدارا حول احد الكواكب , و وصل في نفس الشهر الذي وصل في ه مسبار مارس 2 و مسبار مارس 3 - المسباران الروسيان الي المريخ . و بعد انتهاء فترة اعاصير رملية استطاع ارسال صورا واضحة ل سطح

Keyphrases

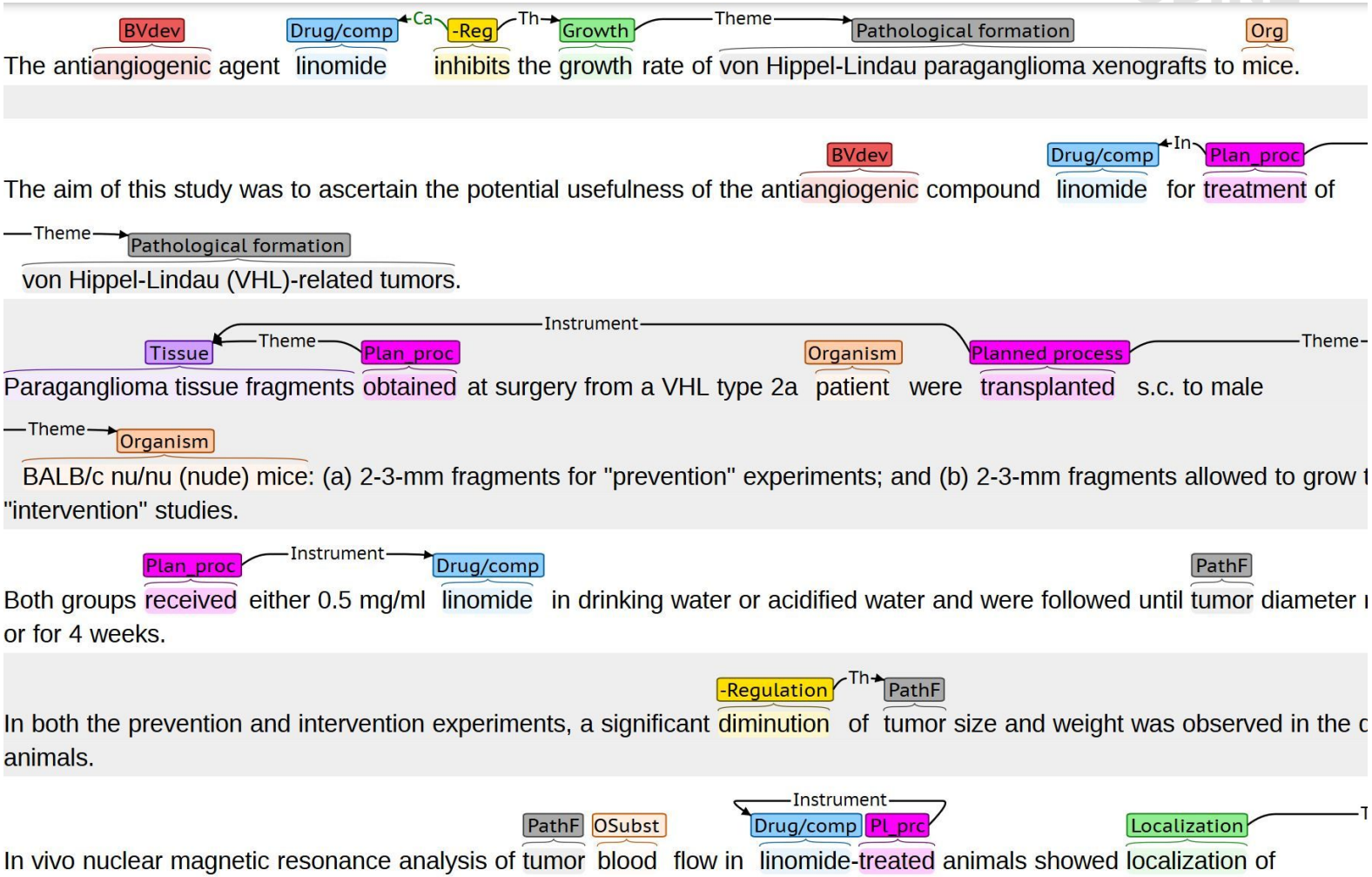
- المريخ
- الارض
- ناسا
- رحلة
- الفضاء
- المسبار
- الشمسية
- مارينر

Biomedical Entity Extraction

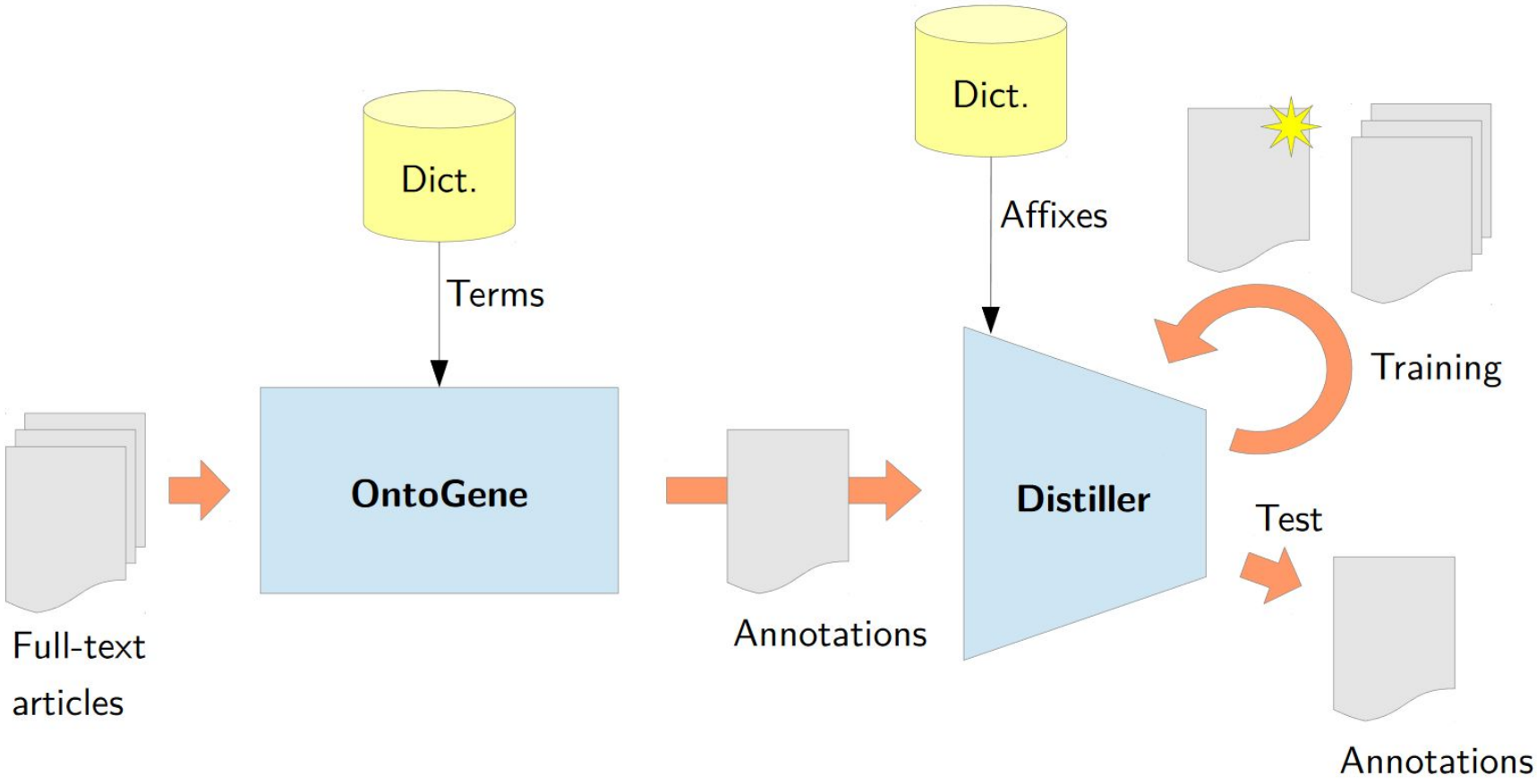
We collaborated with the University of Zurich to create a tool for the annotation of biomedical papers.

- Integration with previous tools (OntoGene)
- Final tool performs ER of specific categories (diseases, proteins, genes…)
- Goal: annotate documents, find relations, provide support for advanced search in corpora (e.g. relation mining)

Example


BVdev
Drug/comp
Ca
-Reg
Th
Growth
Theme
Pathological formation
Org
The antiangiogenic agent linomide inhibits the growth rate of von Hippel-Lindau paraganglioma xenografts to mice.
BVdev
Drug/comp
In
Plan_proc
The aim of this study was to ascertain the potential usefulness of the antiangiogenic compound linomide for treatment of
Theme
Pathological formation
von Hippel-Lindau (VHL)-related tumors.
Instrument
Tissue
Theme
Plan_proc
Organism
Planned process
Theme
Paraganglioma tissue fragments obtained at surgery from a VHL type 2a patient were transplanted s.c. to male
Theme
Organism
BALB/c nu/nu (nude) mice: (a) 2-3-mm fragments for "prevention" experiments; and (b) 2-3-mm fragments allowed to grow t
"intervention" studies.
Plan_proc
Instrument
Drug/comp
PathF
Both groups received either 0.5 mg/ml linomide in drinking water or acidified water and were followed until tumor diameter
or for 4 weeks.
-Regulation
Th
PathF
In both the prevention and intervention experiments, a significant diminution of tumor size and weight was observed in the c
animals.
Instrument
PathF
OSubst
Drug/comp
Pl_prc
Localization
In vivo nuclear magnetic resonance analysis of tumor blood flow in linomide-treated animals showed localization of

Example


Dict.
Terms
Dict.
Affixes
Training
OntoGene
Distiller
Test
Full-text articles
Annotations
Annotations

Biomedical Entity Extraction

- State of the art results on a well-known biomedical scientific corpus
- Needs further work (better disambiguation, speed…)

System	Precision	Recall	F1
OGER	0.32	**0.52**	0.40
OGER+Distiller NN	**0.51**	0.49	**0.50**
OGER+Distiller CRF	0.49	0.29	0.37
MMTx	0.43	0.40	0.42
MGrep	0.48	0.12	0.19
Concept Mapper	0.48	0.34	0.40
cTakes Dictionary Lookup	**0.51**	0.43	0.47
cTakes Fast Lookup	0.41	0.40	0.41
NOBLE Coder	0.44	0.43	0.43

Conclusions

Future work

1. Build a state-of-the-art KE pipeline using the knowledge we gained during these years
2. Integrate Deep Learning tools (see next talk)
3. Use Distiller in "real world" scenarios (e.g. UZH collaboration, other projects)

Distiller

- Try it online: http://ailab.uniud.it/distiller (not the last version though)
- Download it: https://github.com/ailab-uniud/distiller-CORE
- REST endpoint: http://api.ailab.uniud.it (contact us)

Bottom line: the decision of adopting a modular design allowed us to work on very diverse projects

Thank you!

http://ailab.uniud.it/distiller

https://github.com/ailab-uniud/distiller-CORE

http://api.ailab.uniud.it