Predicting Licenses in Changed Source Code

Xiaoyu Liu[1], LiGuo Huang[1], Jidong Ge[2], Vincent Ng[3]

1 Dept. of Computer Science, Southern Methodist University, Dallas, TX

2 State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

3 Human Language Technology Research Institute, University of Texas and Dallas, TX

Emails: {*xiaoyul, lghuang*}*@smu.edu* *gjd@nju.edu.cn* *vince@hlt.utdallas.edu*

After software changes are made …

SMU UT DALLAS The University of Texas at Dallas NANJING UNIVERSITY

Penalties for Software License Violation

Whenever piracy c ilty, they will be responsible for pa ed the software with others, they c y.

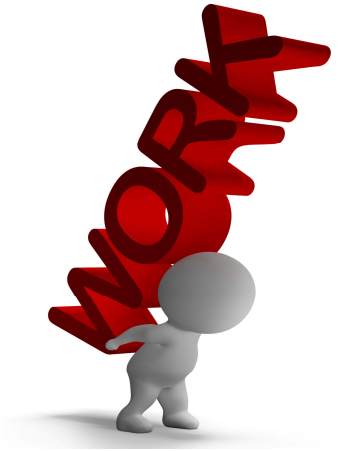
Just like illegally d onsequences may go beyond ju be charged. Depending on the

The legal penaltie They may be required to pay fines as high as $250,000 and may face up to five years in prison. In addition, a permanent felony will be on their record.

Predict license to resolve violation as soon as possible!

Challenges of Software Licensing

- **Time and labor effort consuming**



- **Difficulty: requires lots of experience and expertise**

Software License Prediction

Changed imports in XMLPacker.java
XMLPacker.java was licensed with MPL v1.1

```
22 +import com.sun.org.apache.xalan.internal.xsltc.DOM;
23 +import ke.go.moh.oec.LogEntry;
24 +import java.io.ByteArrayInputStream;
25 +import java.io.IOException;
26 +import java.io.InputStream;
27 +import java.io.StringReader;
28 +import java.io.StringWriter;
29 +import java.text.ParseException;
30 +import java.text.SimpleDateFormat;
31 +import java.util.logging.Level;
32 +import java.util.logging.Logger;
33 +import javax.xml.parsers.DocumentBuilder;
34 +import javax.xml.parsers.DocumentBuilderFactory;
35 +import javax.xml.parsers.ParserConfigurationExcep
36 +import javax.xml.transform.TransformerConfigurati
37 +import javax.xml.transform.TransformerException;
```

***LogEntry* adopts GPL v3+**

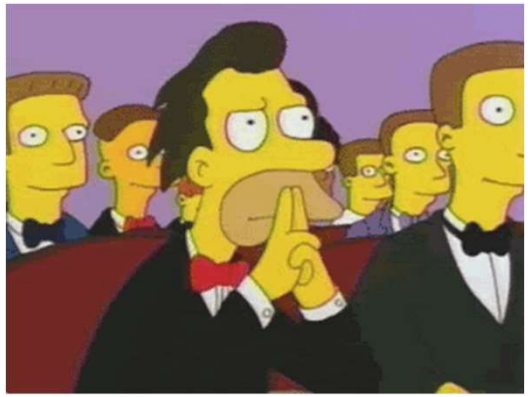
> You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or

(read MPL v1.1 restrictions)

> and other provisions required by the GPL or the LGPL. If you do not delete the provisions above, a recipient may use your version of this file under the terms of any one of the MPL, the GPL or the LGPL.

Goal

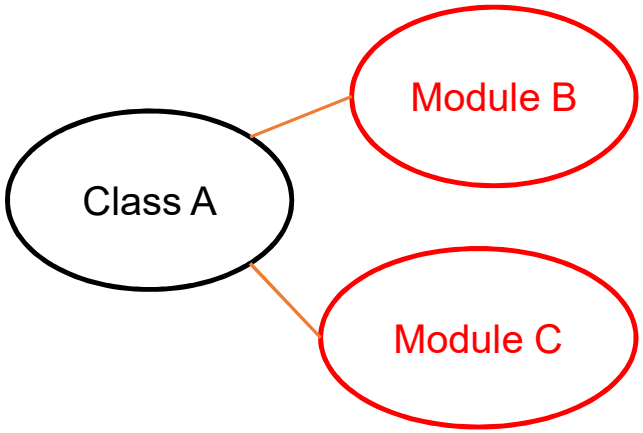
Automatically *predict software license for a* ***changed source code file***



Related Work on File-level License Prediction

- ***Ninka:*** [German et.al., 2010]
  - state-of-the-art, leveraging a set of pre-defined regular expressions built upon these common terms to detect the presence of the license copyrights and terms

- ***Caller-Callee (CC)*:**
  - license based on code imports' licenses

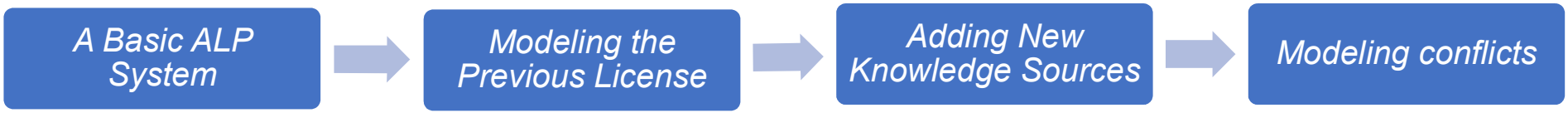


Not taking into account the license compatibility issues of the different license restrictions

Automatic License Predictor (ALP)

- Novel Contributions
  - Propose ALP, the first learning-based approach for license prediction on changed source code file
  - Conflict resolution leverage additional knowledge resources

- Framework



Basic ALP System

changed source code file

"*OpenEMRConnect is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY*"

OpenEMRConnect-distributed; *is-in*; *distributed-the*; *in-hope*; …

code-inline text features

you-not; *may-use*; *not-this*; *use-file*; …

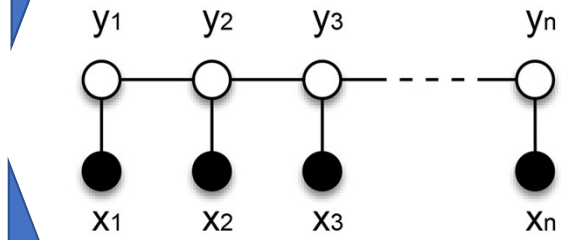
diff features

Logistic Regression

Modeling the Previous License (ALP2)

Previous version

Current version



Conditional Random Field

previous version licensing

current version licensing

Adding New Knowledge Resources

- New resource added!

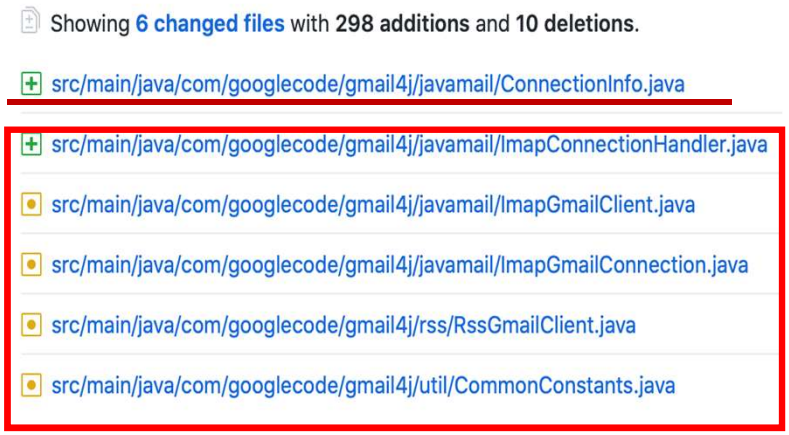
Software documents

LICENSE file

```
MIT License

Copyright (c) 2019 Avery O'Banion

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

Adding New Knowledge Resources

- New resource added!



Co-changed files with
ConnectionInfo.java

Adding New Knowledge Resources

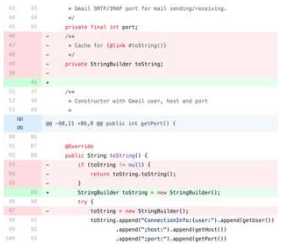
- ALP2+Doc+Co

Software documents

"*retain the above copyright notice*"

retain, retain-above;
the-copyright;
above-copyright;
copyright-notice; …

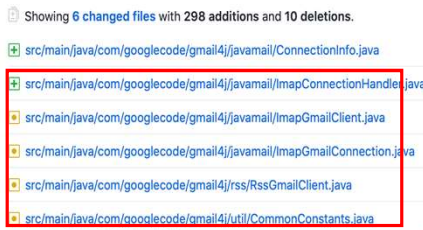
document-text features



Changed source code file

code-inline text features

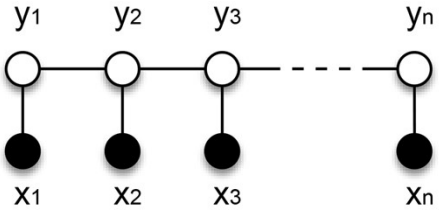
diff features



co-changed source code files

code-inline text features

diff features



Conditional Random Field

SMU

UT DALLAS The University of Texas at Dallas

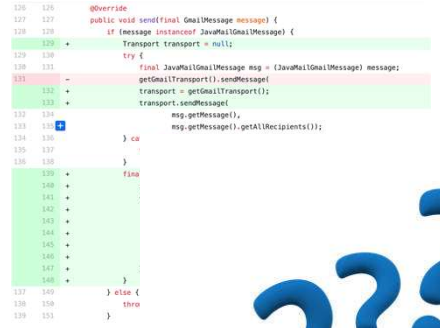
NANJING UNIVERSITY 南京大學

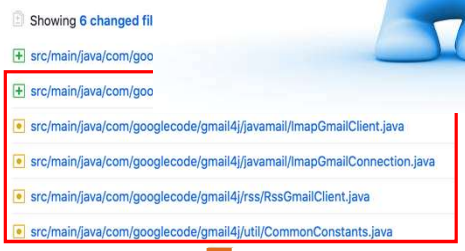
An Example of License Prediction Problem

Changed source code file

Software documents

co-changed source code files





cense A

icense B

License C

SMU

UT DALLAS The University of Texas at Dallas

NANJING UNIVERSITY

Modeling Conflicts (ALP2-Ranker)

- Training

changed source code file

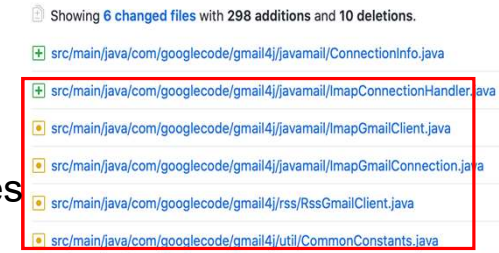
code-inline text features

diff features

software documents

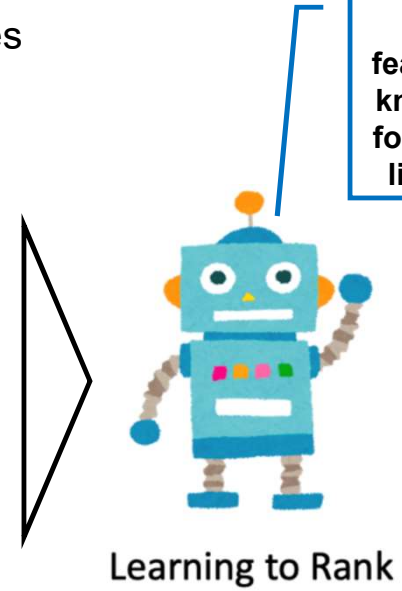
document-text features

co-changed source code files



code-inline text features

diff features

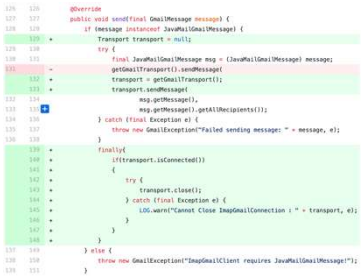


Learning to Rank

Learns the most representative features from each knowledge source for the appropriate license adoption

Resolve Software License Conflicts (ALP2-Ranker)

Changed source code file

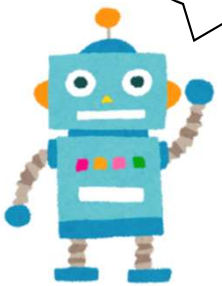


code-inline text features

diff features

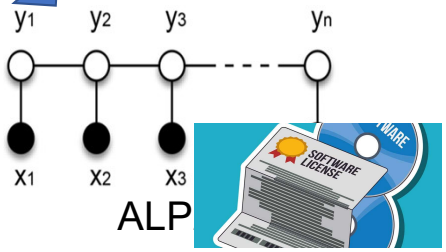
Software documents

document-text features

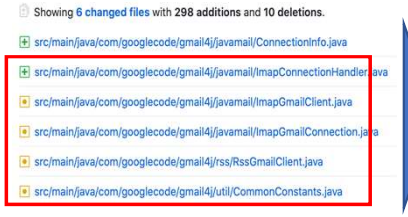


Learning to Rank

1. Software document
2. Changed file
3. Co-changed file



Co-changed source code file



code-inline text features

diff features

SMU

UT DALLAS The University of Texas at Dallas

NANJING UNIVERSITY 南京大學

Empirical Evaluation

- **Datasets: Open Source Java Projects from GitHub**
  - ❖700 Java projects from GitHub

# of systems	700
# of commits	8128
# of changed files	57450

(a) Overall statistics

Procedure of Annotating Licenses for Each File

- We looked at the **terms and conditions for copying, distribution and modifications** of each license of each resource
- **Coders**: two PhD students who have extensive experience in industry as developers
- **Agreement ratio**: 73.7%
- **Disagreements are resolved by open discussion**

Results of Data Annotation

- 25 unique licenses

Licenses	# of changed files
Apache v2	18770 (32.7%)
GPL v2	9458 (16.5%)
GPL v3+	5943 (10.3%)
MIT	3125 (5.4%)
LGPL v3+	2609 (4.5%)
LGPL v2.1+	1542 (2.7%)
BSD	1404 (2.4%)
EPL v1	1276 (2.2%)
Other	4960 (8.6%)
Non-licensed	8363 (14.6%)

(b) Per-license frequencies

Changed file	Software document	Co-changed file
(**Apache v2**) … Subject to the terms and conditions of this License, each Contributor hereby grants to You a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form…	(**GPL v2**) … Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it …	(**LGPL v2.1+)** …linking a "work that uses the Library" with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a "work that uses the library". The executable is therefore covered by this License.. …

Empirical Evaluation

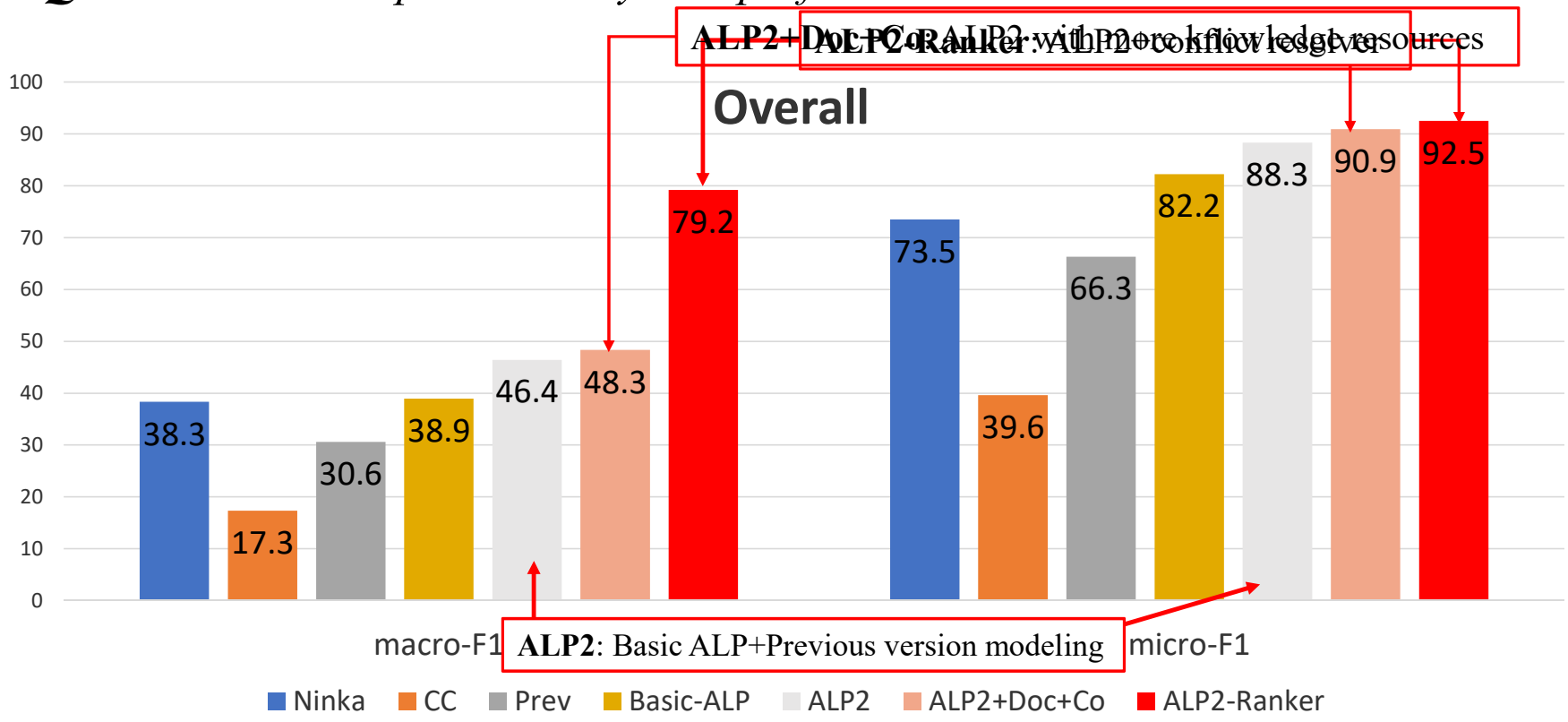
- **Datasets: Open Source Java Projects from GitHub**
  - ❖700 Java projects from GitHub, 25 unique licenses
- **Baseline Systems**
  - ❖ ***Ninka***
  - ❖ ***Caller-Callee (CC)***
  - ❖ ***Previous Version (Prev):*** predicts license to be the same as previous version

Empirical Evaluation

- **Datasets: Open Source Java Projects from GitHub**
  - ❖700 Java projects from GitHub, 25 unique licenses
- **Baseline Systems**
  - ❖ ***Ninka***
  - ❖ ***Caller-Callee (CC)***
  - ❖ ***Previous Version (Prev):*** predicts license to be the same as previous version
- **Metrics**
  - ❖ Macro F1: treat each license as equally important
  - ❖ Micro F1: treat most frequent licenses as more important
  - ❖ Five-Fold Cross Validation

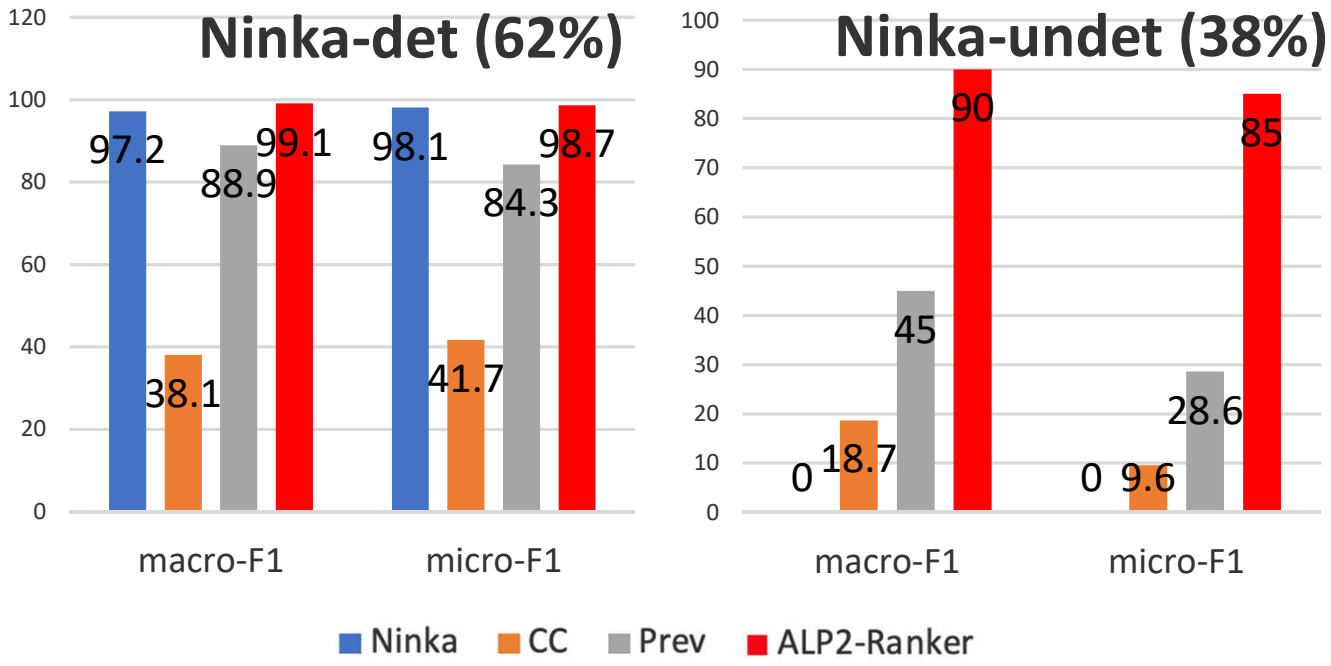
Overall Performance

RQ1. *Which license prediction system performs the best?*

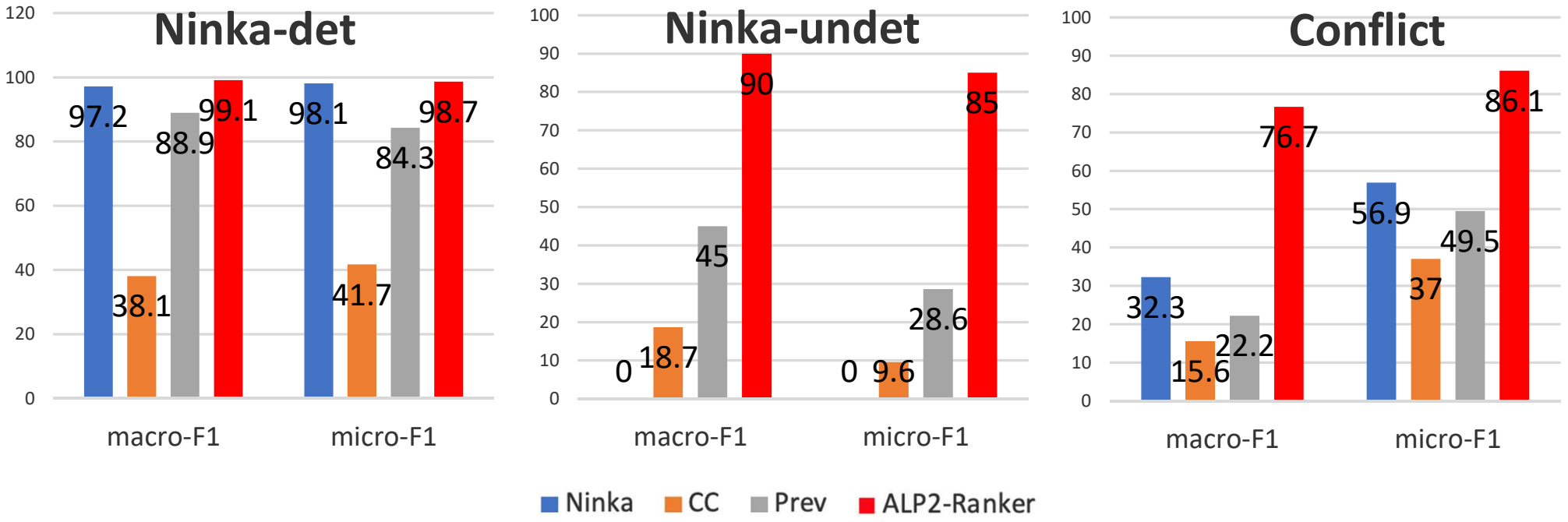


SMU UT DALLAS NANJING UNIVERSITY

RQ2. *How do the systems perform on the easy and difficult tasks?*



RQ2. *How do the systems perform on the easy, difficult, and conflict instances?*



SMU UT DALLAS The University of Texas at Dallas 南京大學 NANJING UNIVERSITY

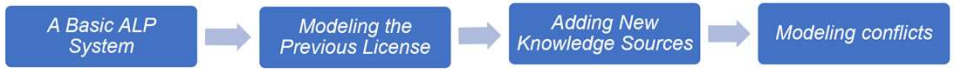
Goal

Automatically** predict software license for a **changed source code file

Automatic License Predictor (ALP)

- Novel Contributions
  - Propose ALP, The first learning-based approach for license prediction on changed source code file
  - Conflict resolution leverage additional knowledge resources

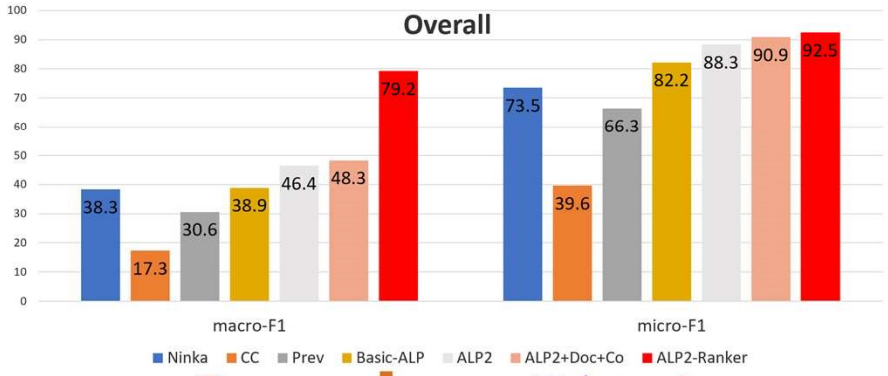
- Framework



Overall Performance

***RQ1.** Which license prediction system performs the best?*

ALP2: Basic ALP+Previous version modeling; **ALP2+Doc+Co**: ALP2+Implicit conflict resolver; **ALP2-Ranker**: ALP2



Predicting Licenses in Changed Source Code

Xiaoyu Liu[1], LiGuo Huang[1], Jidong Ge[2], Vincent Ng[3]

1 Dept. of Computer Science, Southern Methodist University, Dallas, TX
2 State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
3 Human Language Technology Research Institute, University of Texas and Dallas, TX
Emails: {xiaoyul, lghuang}@smu.edu gjd@niu.edu.cn vince@hlt.utdallas.edu