

专业中文简繁转换平台 PCCCP 使用说明

本软件属于 香港城市大学 所有
理念开发、设计与监制：朱纯深
项目主任：朱纯深
技术研究、软件开发：郝天永

一. 简介

专业中文简繁转换平台是一个中文简繁转换工具，开发的目的是为用户提供高质量的中文简繁转换结果。该平台主要具有以下优点：

- 1) 提供高质量简繁转换结果；
- 2) 具有强大的词库训练学习功能；
- 3) 提供自动+高级校订功能；
- 4) 提供用户自定义和本地化词库；
- 5) 使用多种颜色显示不同类型的转换结果；
- 6) 支持多语言、转换结果能保存 Word 文档格式。

本系统提供 6 个主要功能模块，使得普通用户可以快速使用自动转换模式，而专业用户可以使用高级校对功能进行手工校正。个人词典、专业对比评估功能等其他诸多功能的引入，使系统在满足不同用户需求的同时，提供一个专业的中文简繁转换平台。

二. 安装说明

基本硬件需求：

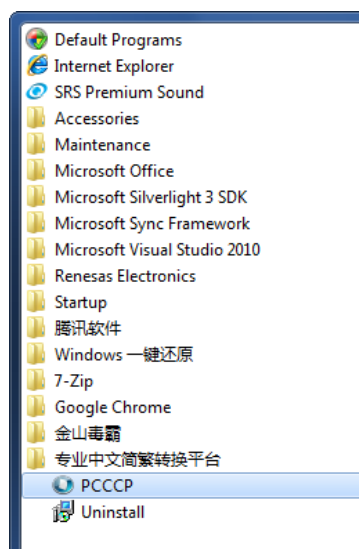
- 1) 个人电脑主机：建议使用 CPU 为 P III 或更高级机种。
- 2) 内存：建议使用 128M 以上 RAM 内存。
- 3) 硬盘：硬盘必需有 60M 以上的空间。
- 4) 显示器：一般 VGA 或 SVGA 显示卡。Windows 色彩显示请设 256 color 或以上，萤幕解析度请设 1024x768 或以上。

支持环境：WinXP 及更高级 Windows 操作系统

所需组件：.Net Framework 3.5 及更高级版本 (Win vista 及 win7 默认含有该组件)

该组件下载地址：<<http://www.microsoft.com/download/en/details.aspx?id=21>>

安装方法：双击【PCCCP_install.msi】即可打开安装引导程序，按照引导程序的指示即可一步步完成安装。安装过程及完成后情况如下页图像所示。



三. 软件注册及语言界面支持

本软件的部分功能免费开放试用，高级功能则不可使用（具体列表请参见第四部分），并在免费版本界面中显示为灰色菜单，不可点击，如图 1 所示。

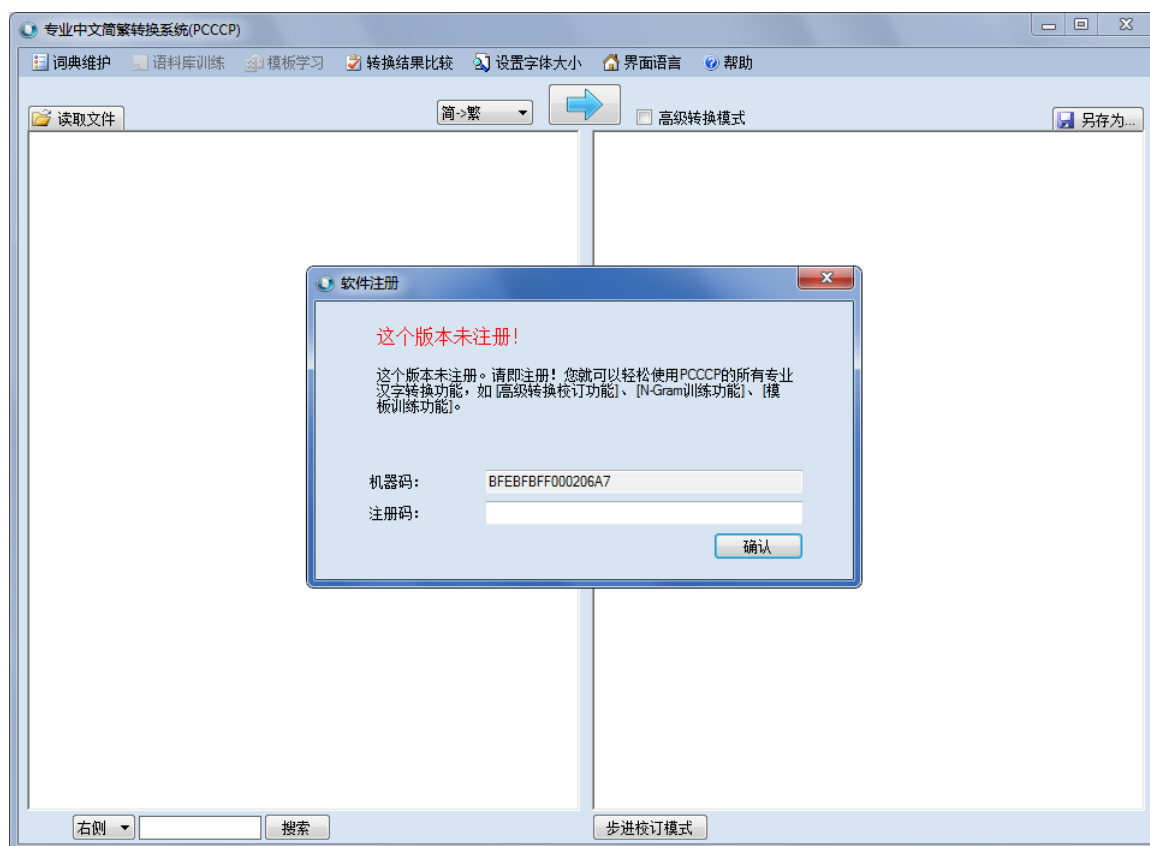


图 1 系统主界面（未注册版本）

如果需要高级功能，必须进行软件注册进而使用。注册方法为：单击【帮助】→【注册】，如图 1 所示，向随后弹出的注册窗口输入与机器码对应的注册码并激活注册，激活后的系统主界面显示如图 2 所示，高级功能显示为正常颜色，并可以点击使用。

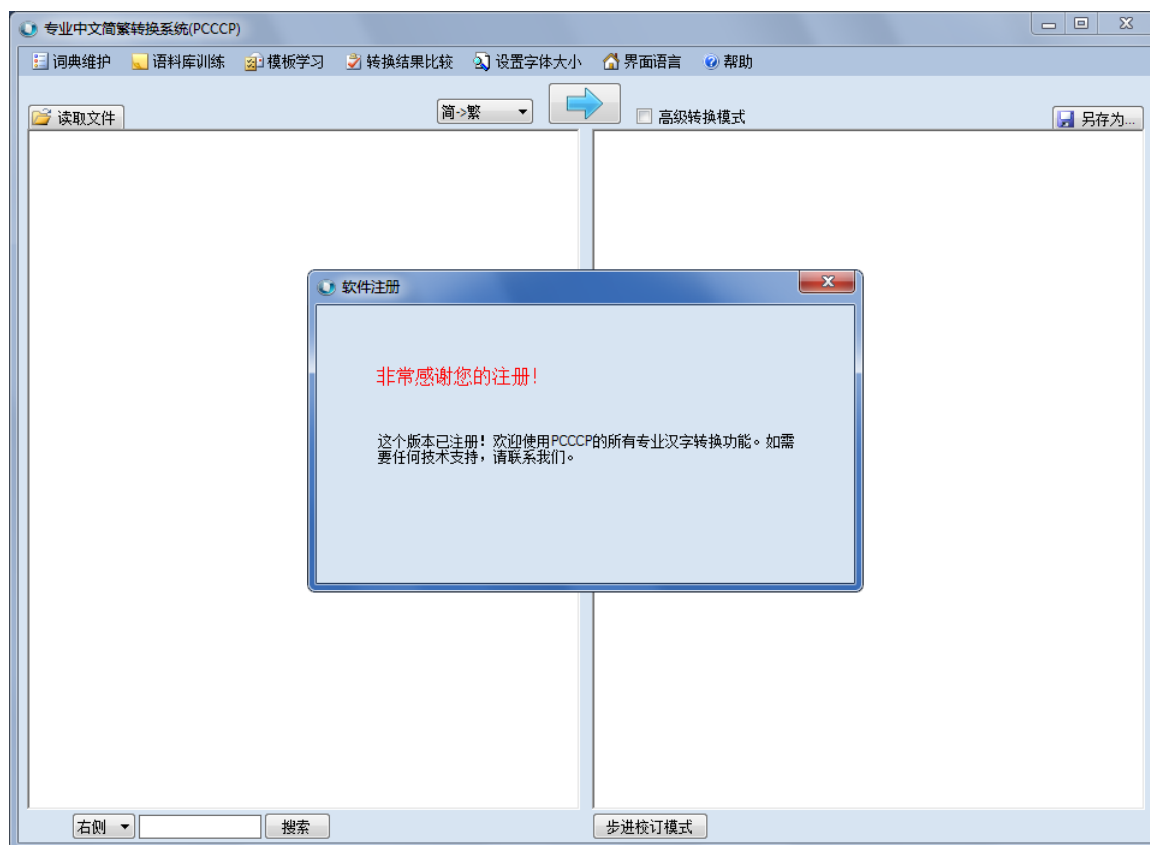


图 2 系统主界面（已注册版本）

本系统现支持英文、简体中文、繁体中文三种语言，因而可以服务不同母语的用户。单击【界面语言】并从随后出现的下拉菜单中选择想要使用的语言类型即可更换页面显示语言，见图 3。

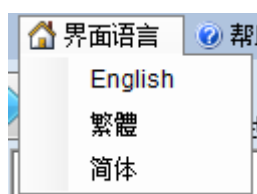


图 3 界面语言设置

四. 主要功能模块

该中文转换系统主要包括以下功能模块：1) 非高级转换模块, 2) 高级手动转换模式, 3) 数据源管理模块, 4) 转换结果对比模块, 5) 语料库训练模块, 6) 模版学习和管理, 7) 提交试用报告。其中 2、5、6 的功能只提供给注册用户使用。

1) 非高级转换模块

使用该系统时, 用户可以直接复制文本 (可以是带格式的 Word 文本) 到左侧区域, 或者加载文档进左侧面板 (点击【读取文件】按钮, 在随后出现的文件选择框中选取.txt 文本文件)。然后, 从下拉菜单中选择【简->繁】或者【繁->简】以进行对应选项的功能, 之后点击选项卡右侧的转换标识按钮进行相应的转换, 转换结果显示在右侧, 默认情况下会使用颜色区分经过转换的字符, 显示结果如图 4、图 5 所示。颜色区分在用户保存文件时自动去除。



图 4 非高级简转繁结果显示示意图



图 5 非高级繁转简结果显示示意图

各种颜色代表的含义如下：绿色--系统自动转换的汉字；灰色--系统自动转换的标点符号。

为方便用户阅读，可单击【设置字体大小】并在随后出现的下拉列表中选择字体大小。或者将光标置于左侧，然后按住 ctrl 键，并同时滑动鼠标滚轮放大左侧字体，右侧可通过同样的方式调整字体显示大小。

对于图 4 所示的非高级转换结果，用户可以进行手动校验，用鼠标选中待修改的字，系统会自动定位到左侧与之对应的字，并在右侧弹出一个修改框，可直接在修改框中输入校正后的字，或者从下拉列表中选择合适的字，然后单击【确定】按钮完成对该字的修改。修改窗口如图 6 所示。



图 6 结果修改窗口

图 6 右侧的修改小窗口中还提供了两个修改选项：更新所有相同情况、先搜索再更新。如果选中【改正所有相同情况】，则单击【确定】按钮后会修改所有与当前选中的字相同的字的处理结果。如果选中【先搜索再更新】，则小窗口显示如图 7 所示。先点击【搜索】定位到特定的字，再单击【替换】可对该字的转换结果进行修改。

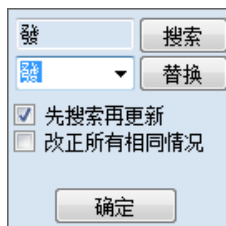


图 7 先搜索再更新

所有修改完毕后，单击右侧【另存为】按钮，可将转换结果保存到指定位置。

2) 高级手动校正模块

该模块允许用户选择需要的数据源进行特定任务的转换，比如地区性词语转换。转换结果也进一步使用不同的颜色区分不同的操作。该模块更允许用户批量校正相似的转换错误，并可以把校正后的内容加入到个人词典使得将来的转换能自动避免该错误。有了该模块的帮助，用户可以快速、高效地以手动方式校正机器自动转换出现的错误。加之灵活的数据源选择，该模块更适合专业人士进行快速、批量校正。

模块的主要使用方法如下：点击【高级转换模式】。☒高级转换模式 随后点击处弹出一个包含多个数据源的窗口，如图 8 所示，用户可以根据需要选择其中的单个或者多个数据源，然后点击转换按钮再次对左侧按钮进行转换。

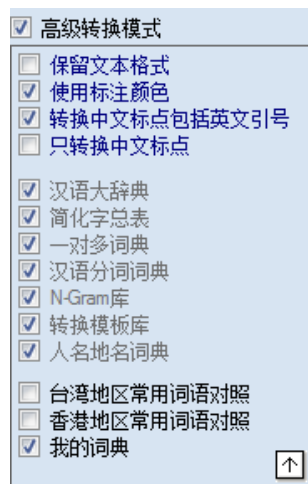


图 8 多数据源选择窗口

转换结果（如图 9 所示）已使用颜色进行区分，绿色为机器自动一对一字符转换结果（该部分准确率极高，高于 99.8%，一般不需要用户确认），黄色为机器自动一对多字符转换结果（该部分准确率为 91.5%，需要用户校正），灰色为机器自动标点符号转换结果。如果用户对转换结果进行校正，校正后的字符以蓝色显示。



图 9 使用高级简转繁模式得到的转换结果显示



图 10 使用高级繁转简模式得到的转换结果显示

用户可以对转换结果进行快速校正，比如用户发现一处错误，可以点击该字符或者单词，该内容附近即弹出一个校对窗口（见图 11），用户可以从下拉菜单中选择正确的字符（如有）或者直接手工填写。同时，该窗口包含 3 个选项：加入个人词典、改正所有相同情况、先搜索再更新。后两个按钮的功能同普通模式，修改结果用蓝色表示；加入个人词典表示用户的当次校正及正确结果将会自动加入用户词典防止下次出现相同的转换错误。当用户完成校正后，可以使用取消颜色功能取消特定颜色，或者取消全部颜色以得到最后结果。

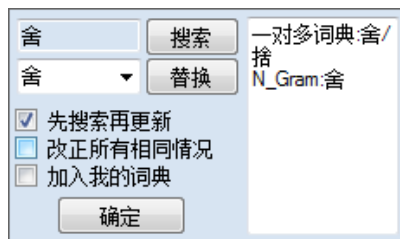


图 11 高级转换模式下的校对窗口

高级转换模式下还提供步进校订模式，单击【步进校订模式】[步进校订模式](#)按钮，系统会自动定位到可能需要手动校订的字或词组供用户校订，校订完毕后自动定位到下一个可能需要校订的字或词组。如果想中断步进校订状态，只需在右侧空白处单击鼠标即可。修改完毕后，

仍然通过右上侧【另存为】按钮保存当前转换结果到指定位置。

2.1) 标点符号转换

使用该系统进行标点符号时，用户可以选择 ☐ 只转换中文标点 即仅仅转换中文标点符号，或者可以选择 ☒ 转换中文标点包括英文引号 来转换中文标点符号和英文标点符号，需要说明的是，第 2 种转换模式下，系统可以进行智能判断，即英文标点符号内部如果均是英文字符的情况下，系统认为该文本是英文引文，因为不需要转换标点符号。如图 12 所示，在第 2 种转换模式下，第一句为英文引号和英文内容，因此不需要转换；第二句使用中文引号和英文内容，因此被转换；第三句为英文引号但内含中文字符，因此被转换；第四句为中文引号和中文内容，因此被转换。

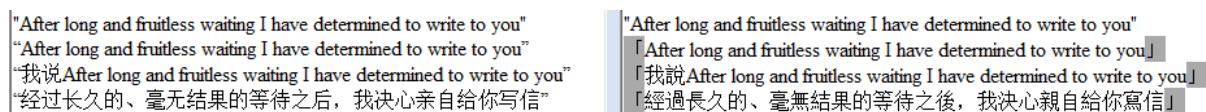


图 12 转换中文标点包括英文引号

转换后的结果自动用灰色着色，以清楚显示，用户可以点击右下角对应的色块来去除该颜色。

2.2) 右键菜单

使用该系统进行校订时，用户可在待转内容（左侧框）和已转内容（右侧框）中任意地方点击鼠标右键，激活右键菜单，如图 13 所示。该多语言菜单包含“复制”、“粘贴”、“剪切”、“删除”、“撤销”、“全选”等常用功能，需要说明的是：在校订过程中可以使用“撤销”操作来恢复上一个操作状态，例如，撤销上一个操作中对一个字的着色或修改。

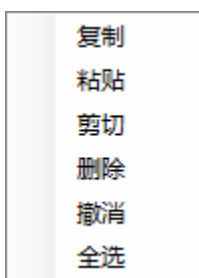


图 13 左侧、右侧框都可以激活的右键功能菜单

3) 数据源管理模块

本系统提供一个用户可以浏览、管理所有数据源的功能模块，用户可以使用插入、删除、保存、另存、搜索等诸多功能。同时，一些字符由于非常相似，比如“麪”和“麪”，系统提供了放大功能（点击想要放大的字符即可），方便用户浏览字符间的细微不同。

单击【词典管理】→【词典管理】，打开词典管理界面，如图 14 所示。

单击显示窗口中的任意词条，系统自动将该词条对应的转换情况显示在窗体下侧，如图 16 所示。

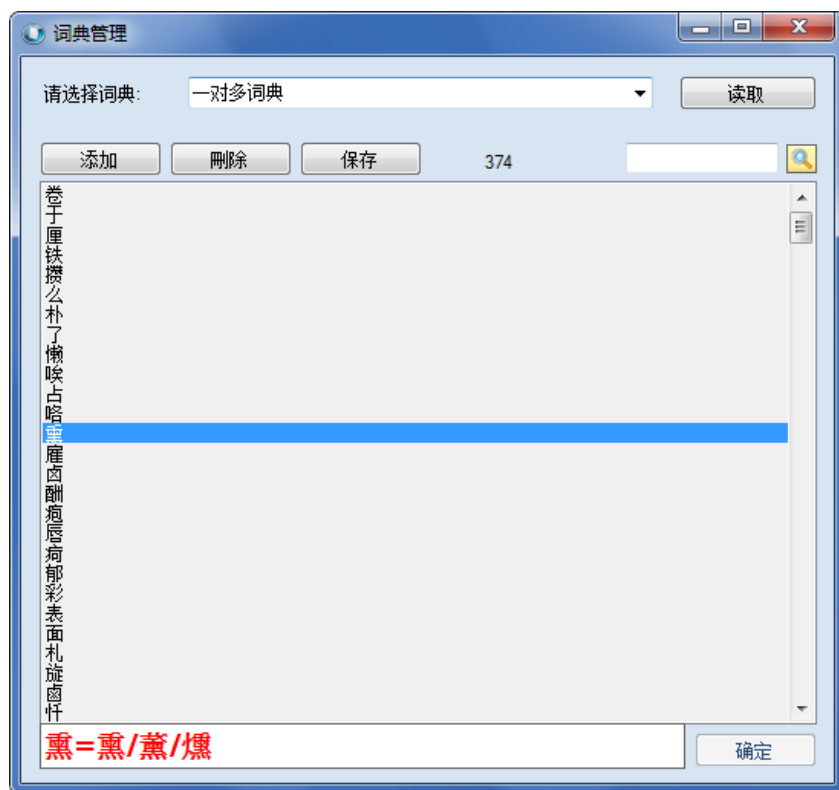


图 16 词条显示示意图

如果需要修改该词条的内容，直接修改下侧的红色字体即可，如果需要添加新的词条，单击【添加】按钮，系统弹出操作说明提示框，如图 17 所示。

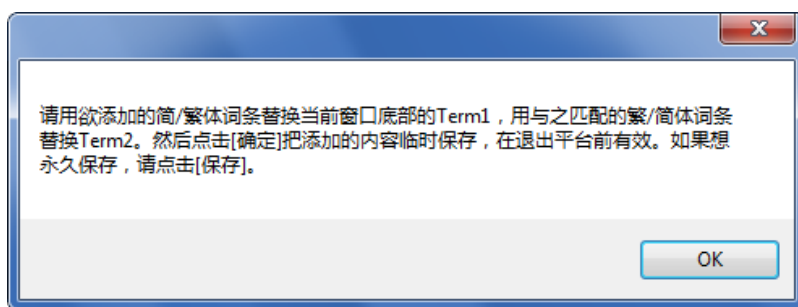


图 17 添加词条操作说明示意框

单击【确定】之后，窗口显示如图 18 所示。修改下侧 Term1 与 Term2 的内容即可完成词条插入操作。

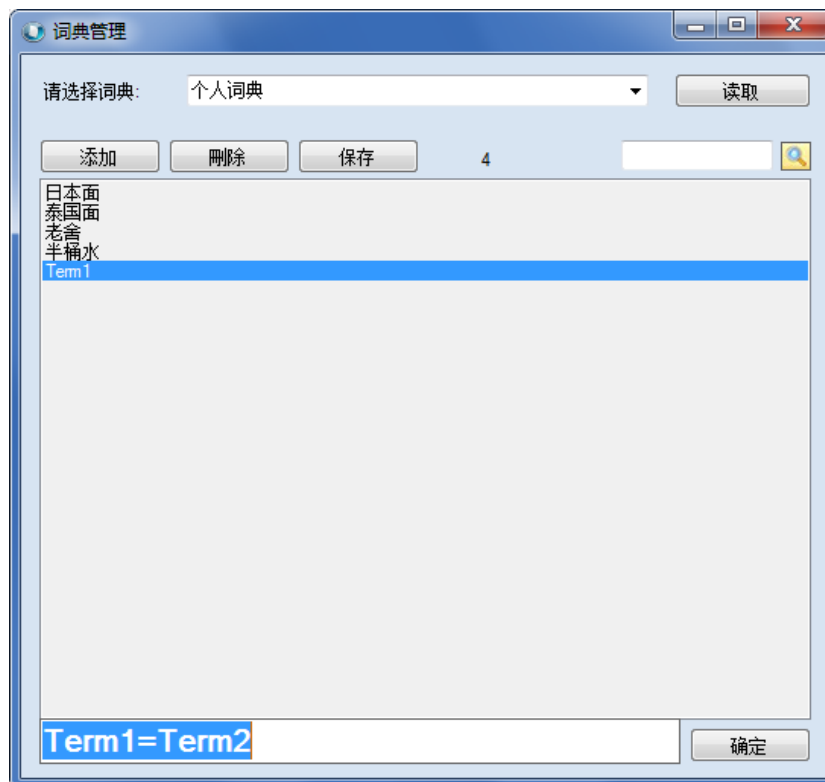


图 18 新加入的词条显示示意图

如需删除词条，先选中该词条，然后单击【删除】按钮。所有的增删改操作完成后，单击【保存】按钮，将之前所做的各种修改保存到物理磁盘。

4) 转换结果对比模块

本系统更进一步提供转换结果对比功能和统计报告功能，用户可以使用该模块来对比不同转换工具对同一文本的转换结果。如果用户已经有正确的转换结果，更可以使用该正确结果跟任意一个工具生成的转换结果进行对比，以评估该转换结果的转换精度。

使用时选择【转换结果比较】打开转换结果比较窗口，然后分别读取文件到左侧和右侧面板，之后单击【比较】按钮，系统即可进行比较并输出结果，如图 19 所示。

在对比功能中，不同转换结果的细微不同都使用颜色进行区分，左侧统一使用黄色，右侧使用绿色，使得用户可以快速辨认这些不同。左下角的报告提供了这些不同的具体信息，包括字符的具体位置。在生成的报告中，该模块提供了详尽的结果，主要包括：总字符数统计，不同的字符数统计，不同字符数比例统计，整体转换精度统计，一对多字符数统计，不正确的一对多字符转换统计，一对多转换精度统计，该统计以左侧作为标准。

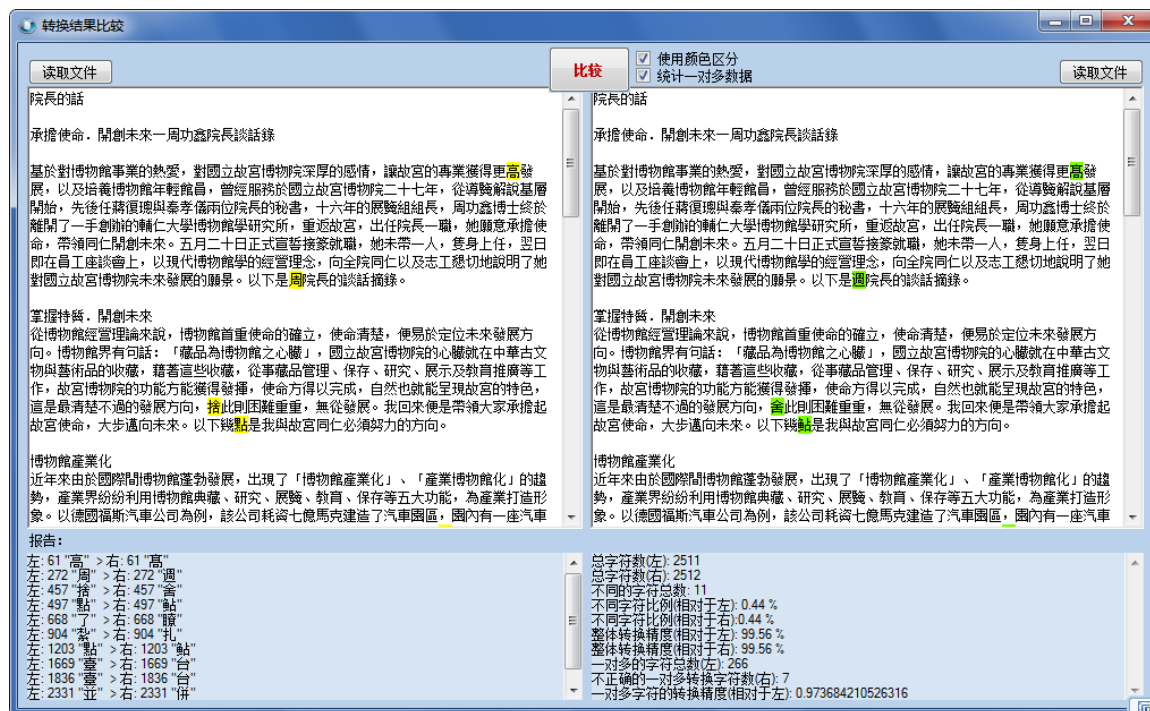


图 19 转换结果比较窗口示意图

5) 语料库训练模块

对于专业用户，本系统允许用户从自己的语料库中批量训练 N-Gram。该训练语料需为繁体中文因为该 N-Gram 主要用于由简体到繁体的转换。单击【语料库训练】打开语料库训练窗口，如图 20 所示。

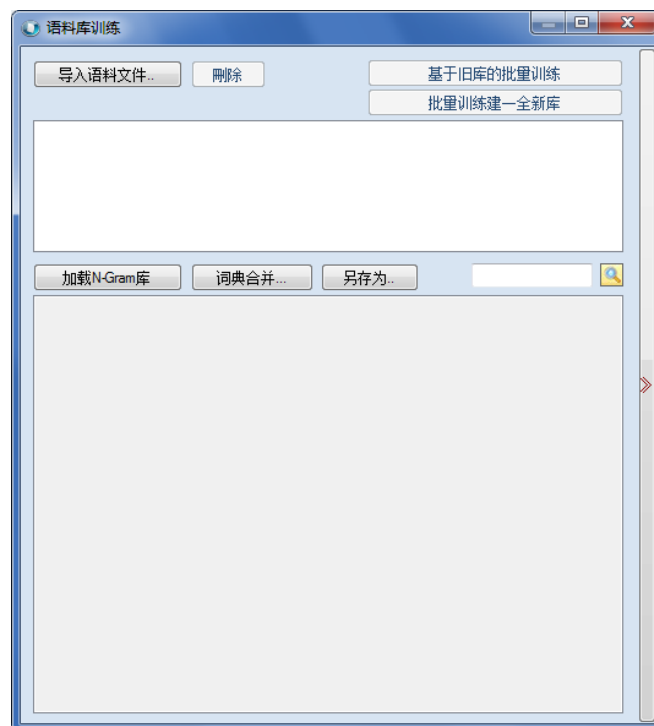


图 20 语料库训练窗口

首先单击【导入语料文件...】打开文件选择窗口选择繁体文本语料文件，一次只能选取一个文件，如需同时训练多个文件，需反复执行多次导入语料文件操作，如果导入了不想要的语料，可先选中该语料，然后单击【删除】按钮移除该语料。语料选择完毕后，对当前导入的所有语料，有两种训练方式，其一是【基于旧库的批量训练】，其二是【批量训练建一全新库】。第一种方式在已有的 N-Gram 库的基础上进行训练，速度较慢，需要等待较长时间才能得到结果，第二种方式速度相对较快。训练完毕后，训练结果输出到窗口下侧，如图 21 所示。

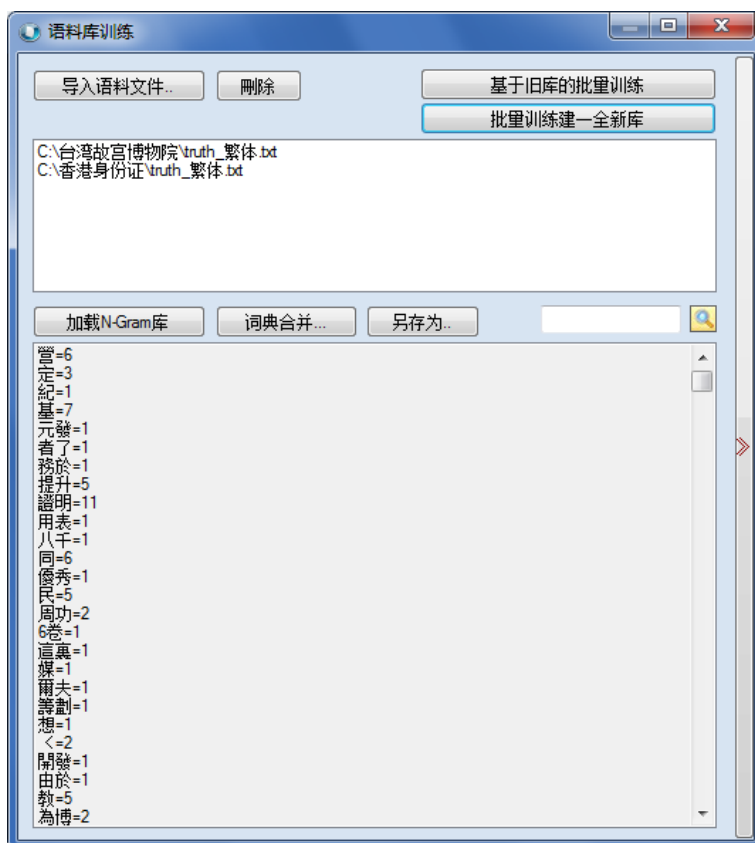


图 21 批量训练输出结果示意图

生成的结果形如“營=6”，意思是“營”字在当前的库中出现了 6 次。用户可以管理获取的 N-Grams 或者合并不同的训练结果，单击【词典合并】按钮，可选择已有的词典，单击【另存为】按钮，可以保存当前训练结果为 dat 文件。

同时，系统提供了一个对训练结果的测试对比功能（在右侧窗口，见图 22，需要单击折叠条打开右侧窗口）。用户可以选择最大逆向匹配、最大逆向匹配+N-Gram、最大逆向匹配+N-Gram + 模版、微软 Office。每种方法对文本进行处理的结果和相应的运行时间显示在窗口右下，有助于用户全面分析测试结果和程序运行效率。

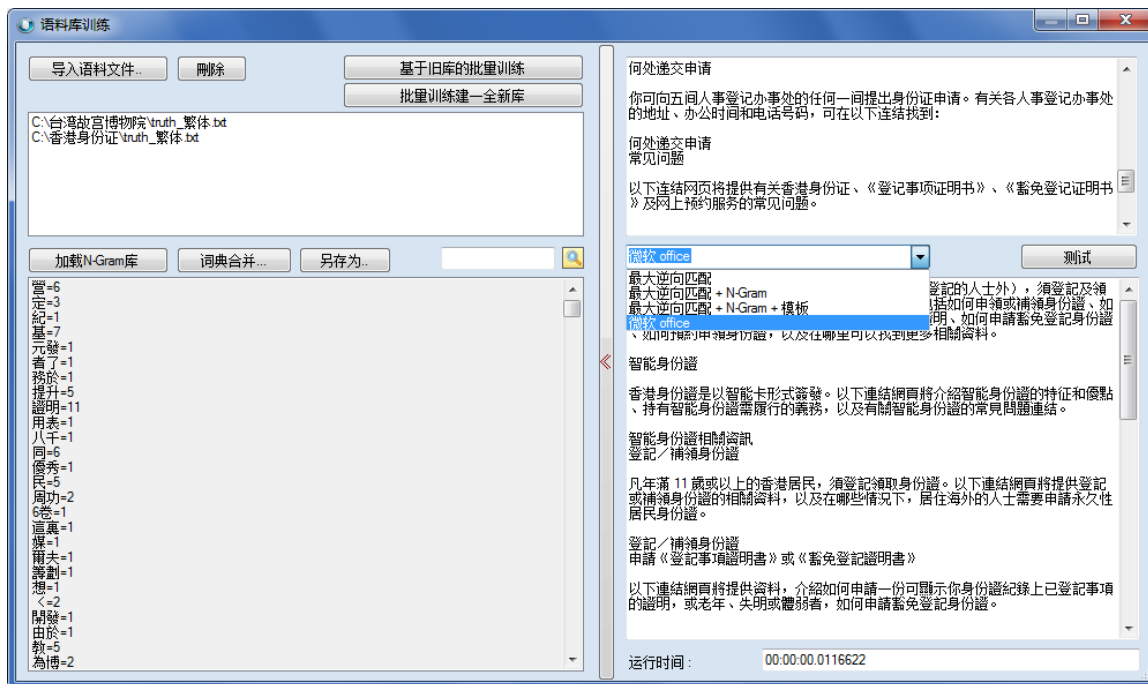


图 22 训练结果测试对比显示示意图

6) 模版学习模块

本系统提供一个模版学习模块，使用户可以自己训练学习模版用于特定领域的文本转换中。单击【模板学习】，打开模板学习窗口，如图 23 所示。

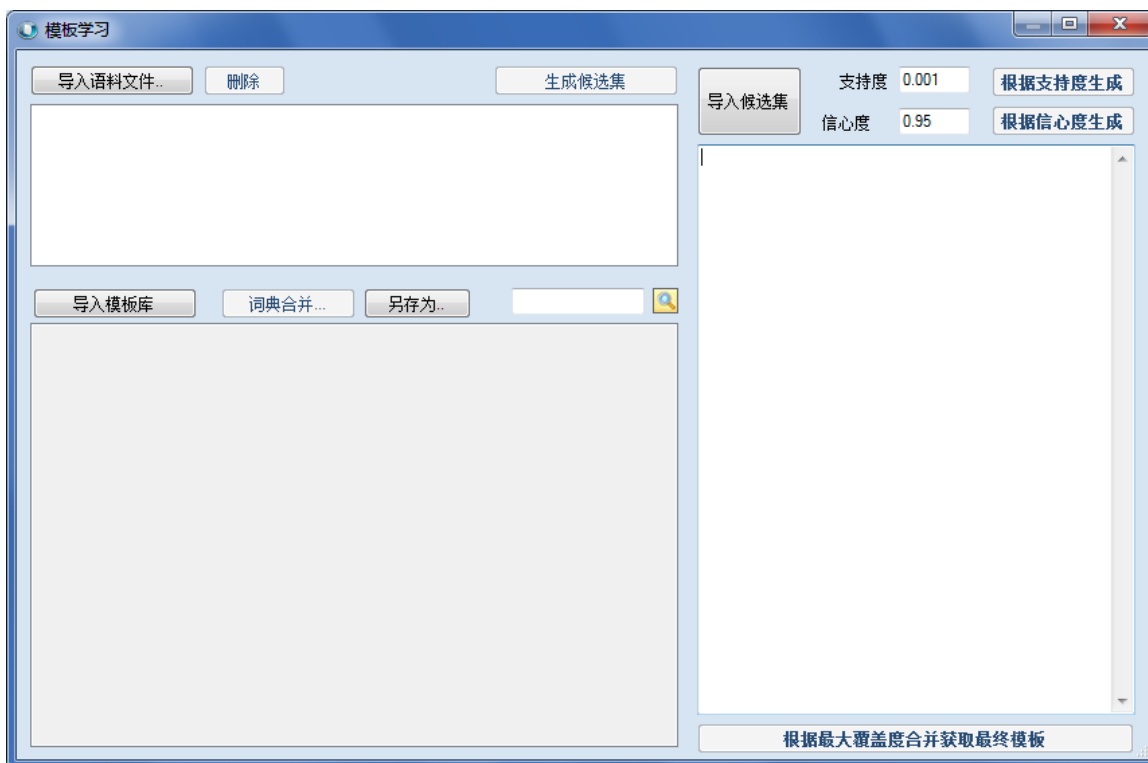


图 23 模板学习窗口

首先单击【导入语料文件...】按钮导入待训练的语料文件，但后单击【生成候选集】按钮，如果训练语料较大，则此过程会比较漫长，请耐心等待，训练完毕后，系统会将结果输出到左下侧区域，如图 24 所示。单击【导入模板库】，系统显示如图 25 所示

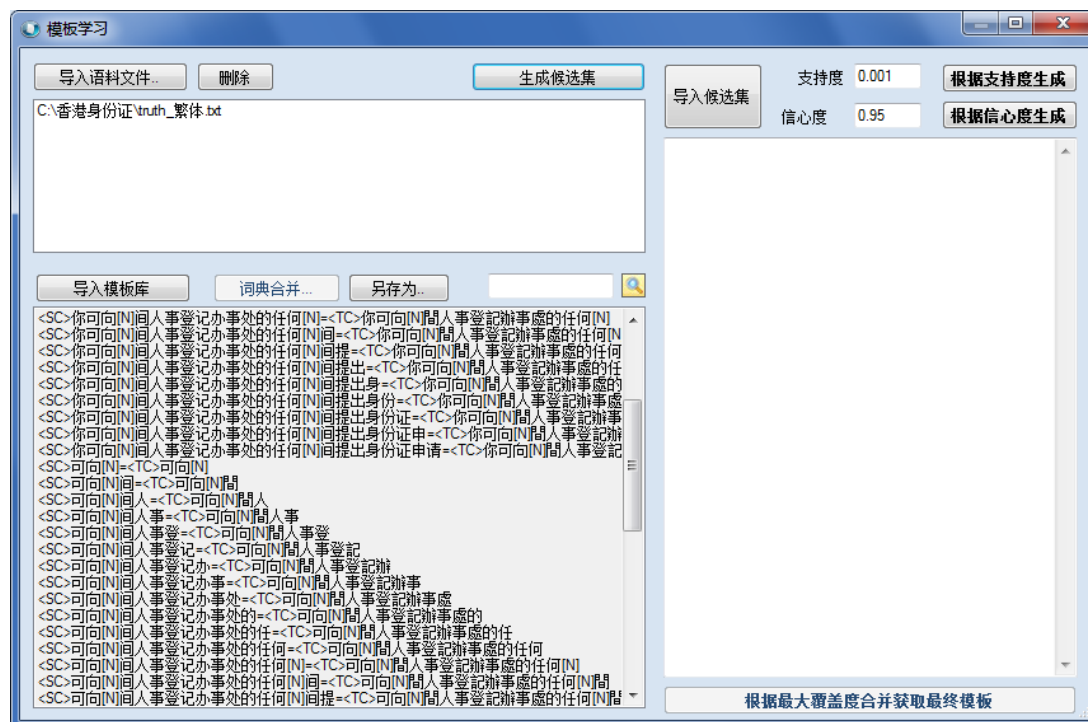


图 24 模板学习输出结果示意图

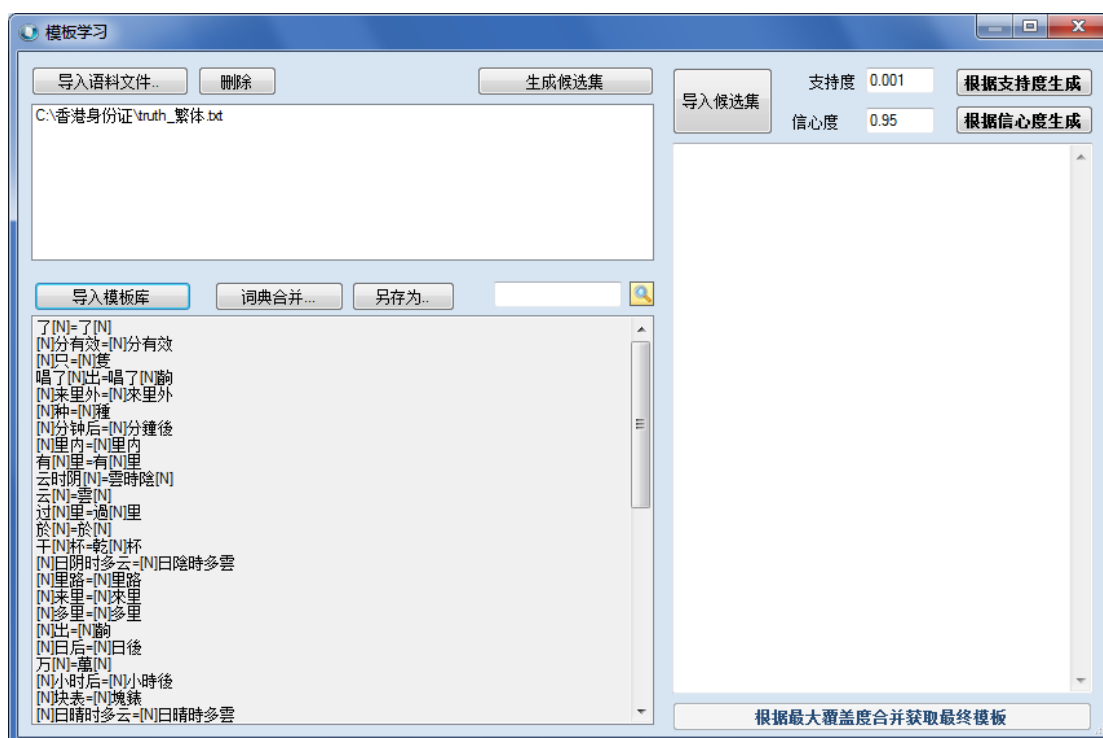


图 25 导入模板库后的显示示意图

单击【另存为】按钮，可将当前训练结果存储为 dat 文件。单击窗口右侧的【导入候选集】按钮，选择 dat 文件并打开，见图 26，然后可设置支持度和信心度来进行基于支持度和信心度的候选集过滤，得到的合适的模版，可以通过“根据最大覆盖度合并获取最终模版”来合并并获取最终模版。该训练和学习过程允许用户运行和操作每个步骤，并具体控制和调整参数，因而更加灵活。

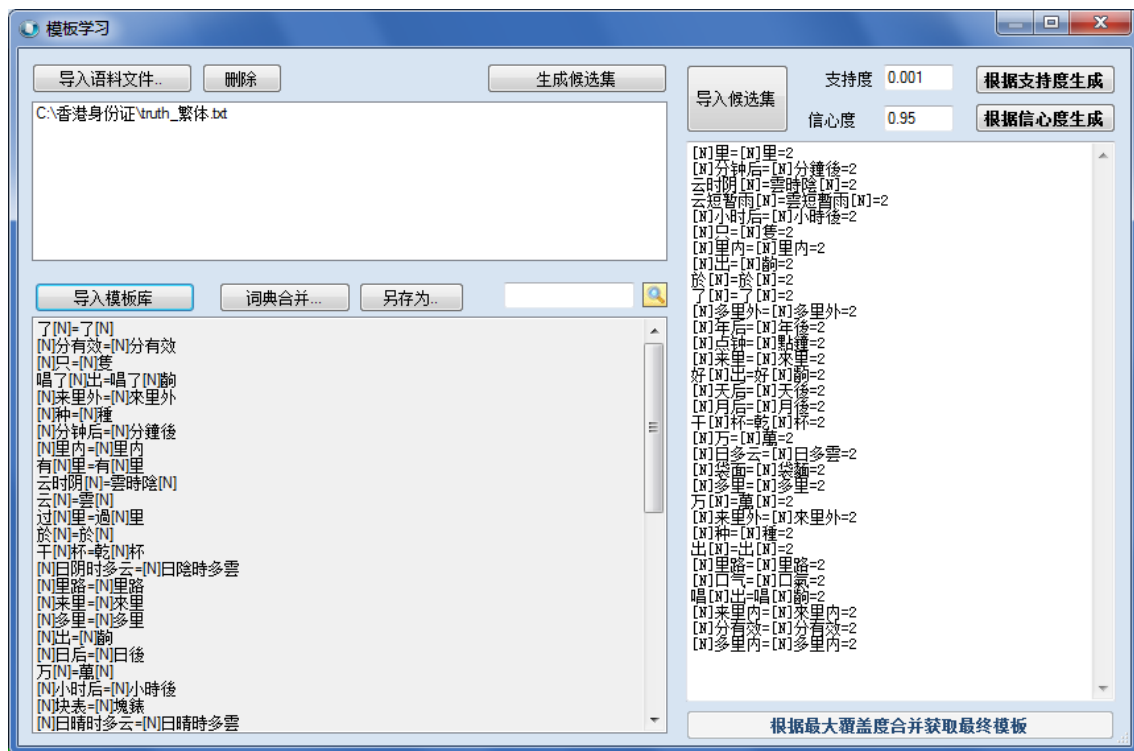
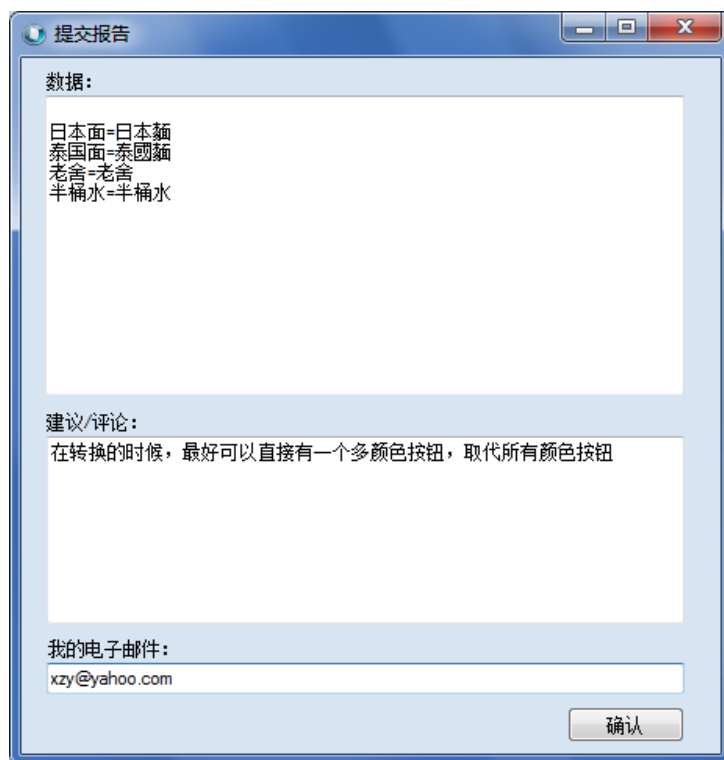


图 26 用户自选参数进行候选集过滤

7) 提交试用报告

为更好的改进本软件，我们鼓励用户提交各种类型的试用报告，可以是建议或者评论，用户的使用数据也同时提交，如需提交试用报告，请单击【帮助】→【查看并提交我的试用报告】，向随后的弹出窗口中，输入建议/评论和个人的邮件地址并提交，相应的界面显示如图 27 所示。



提交报告

数据:

日本面=日本麵
泰国面=泰國麵
老舍=老舍
半桶水=半桶水

建议/评论:

在转换的时候，最好可以直接有一个多颜色按钮，取代所有颜色按钮

我的电子邮件:

xzy@yahoo.com

确认

图 27 用户填写并提交试用报告